

Integrating ethical principles into AI systems: Practical implementation and societal implications

Franziska Marie Poszler

Vollständiger Abdruck der von der TUM School of Management der Technischen Universität
München zur Erlangung einer

Doktorin der Philosophie (Dr. phil.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Svetlana Ikonnikova

Prüfende der Dissertation:

1. Prof. Dr. Christoph Lütge
2. Prof. Dr. Sven Nyholm

Die Dissertation wurde am 15.04.2024 bei der Technischen Universität München eingereicht
und durch die TUM School of Management am 23.07.2024 angenommen.

The duty of man is the same in respect to his own nature as in respect to the nature of all other things, namely not to follow but to amend it.

(Mill, 1874¹)

¹ Mill, J. S. (1874). *Nature – The utility of religion and theism*. Longmans.

Acknowledgments

Through the course of this dissertation, I have had the privilege of engaging with numerous remarkable individuals. They provided invaluable support in shaping this academic work and played a vital role in my personal growth. While I cannot list everyone here, I would like to extend my acknowledgments to those whose impact was exceptionally profound.

First, I would like to express my sincere gratitude to my doctoral supervisor, Prof. Dr. Christoph Lütge. I am extremely thankful for how much freedom you granted me in my research activities and for how much trust you placed in my work. Thank you for your friendliness, empowering me in my academic career, and for introducing me to unique opportunities. I also would like to thank Prof. Dr. Christoph Lütge and Dr. Caitlin Corrigan for providing me with funding throughout this dissertation by hiring me as a research associate at the Chair of Business Ethics and the Institute for Ethics in AI (IEAI). Being part of the launch and development of the IEAI from day one was inspiring and insightful.

In the same vein, I would like to thank a few additional senior scholars who have shaped my academic career and/or acted as co-authors in joint research projects. I am immensely grateful to Prof. Dr. Gari Walkowitz for mentoring me over the past few years and for providing me with valuable advice on my scientific studies, and essential career steps in the future. I really enjoyed our meetings and learned a lot from them. Furthermore, I would like to thank Prof. Dr.-Ing. Johannes Betz, who initiated the ANDRE project and who supported me exceptionally during application processes. Moreover, I would like to express special thanks to Prof. Dr. Sven Nyholm for your thoughtful feedback and for taking the time to act as my secondary examiner for this doctoral dissertation. Similarly, I would like to thank Prof. Dr. Svetlana Ikonnikova for chairing my dissertation's examination board. I also would like to thank Prof. Dr. Isabell Welp and Prof. Dr. Maria Strobel for kindly supporting me when I decided to pivot my academic career towards AI ethics in 2018.

I would like to thank ALL my colleagues and students I met and worked with during my time at TUM, from the Chair of Business Ethics, the IEAI, and the Chair for Strategy and Organization. I want to mention Ann-Carolin Ritter, with whom I stuck together through any tasks and situations. I couldn't have imagined a more collegial fellow to have started the academic career with. I am very grateful for my project partner, Maximilian Geisslinger. Thanks for being

so smart, pragmatic, thoughtful, ambitious, and just a lot of fun to work with! We didn't manage to circle the Arc de Triumph with a self-driving vehicle, but I am very proud of the results we achieved in the ANDRE project. Furthermore, I would like to personally mention Dr. Anna Schulte Steinberg, who became a friend over the years as we sat side by side to each other in the office. I want to say thanks to Dr. Raphael Max for his kindness and readiness to help. Thanks to my long-term office roommates, Dr. Nicholas Folger and Dr. Marianne Thejls Ziegler. I really enjoyed your company and the insightful and amusing talks we had. Additional credits go out to the people who I believe are the backbone of the chairs/institute I worked at, especially: Anastasia Aritzi, Christina Daman, Manuela Fuchs, Tamara Partetzke, Doris Meier, Julia Howard, Eva Pongracz, and Petra Langhanki. Thank you so much for helping with many administrative, organizational, team-building, accounting, or outreach tasks and lighting up the academic atmosphere with your energy.

Finally, I want to express my deepest gratitude to my family and friends. First and foremost, I would like to thank my parents, Silvia and Laszlo: Any words I put down here won't adequately convey the depth of my appreciation for you and all you've done for me. Thank you for your endless love, support, and encouragement and for providing our family with a home where I always feel at ease and welcome. I am also beyond grateful to Sandi, the best brother I can imagine. Thank you for being so caring, supportive, dedicated, modest, dynamic, intelligent, and silly at the same time. You have inspired me throughout my life. Additionally, I am so happy to have my loving grandparents, Siegrun and Manni, by my side. Thank you for the close relationship we share, for always being proud and interested in what I do, and for making sure, I am okay despite any professional commitments. I would also like to mention my grandparents, Livia and Laci, who, unfortunately, cannot witness the completion of my doctorate anymore but who always showed me unconditional love and handed down their enthusiasm for scientific inquiry and philosophy to me. I am forever thankful for my close friends and family-in-law who supported me during this entire journey and who bring me so much joy in life. I want to especially call out Henni for being the biggest cheerleader and such a loyal and genuine friend: Thanks for celebrating every important milestone with me. Lastly, I want to thank Laurenz for being more critical than any journal reviewer 2 and more time-conscious than any strict deadline. Joking aside, thank you for your understanding, for giving me space to pursue my dream profession, and for being such a supporting, fun, and faithful companion. Thanks to your sweet surprise in December 2022, in my memory, I will always associate the time of my dissertation with our promise to spend the rest of our lives together.

Abstract

AI systems are increasingly deployed in contexts where computed decisions have ethical implications in that they touch upon vital human values. For instance, self-driving vehicles must decide how much distance to keep from other traffic participants or, in extreme cases, who to crash into when an accident is inevitable. Decisions like these have palpable effects on values such as ‘safety’ and ‘fairness’, essentially determining what constitutes a fair distribution of risk among individuals in traffic. Companies need to program these decisions in advance to ensure that AI systems can make explicit use of ethical considerations in their decision-making, thereby creating artificial moral agents. However, the question of how to effectively operationalize and instantiate high-level values within AI systems still needs to be resolved in the field of machine ethics. Therefore, this doctoral dissertation investigates how to integrate ethical principles into AI systems, specifically focusing on self-driving vehicles. It identifies the activities practitioners can undertake in this endeavor and examines corresponding societal implications.

To do so, this dissertation consists of three self-contained essays. Essay I relies on qualitative interviews with experts in philosophy, AI, and cognitive sciences to examine arguments for and against computational ethics and artificial moral agents. It also summarizes indicated recommendations for the technological design and development processes of companies, as well as governance measures for the industry. Essay II conducts a systematic review and integration of the autonomous driving ethics literature to evaluate the applicability of various ethical theories to the decision-making of self-driving vehicles. Essay III establishes system requirements and a precise ethical decision-making model for self-driving vehicles facing hazardous situations in traffic. Overall, this dissertation aims to lay the groundwork for turning AI ethics into practice by showcasing how to develop ‘ethical’ artificial moral agents, using self-driving vehicles as an example.

Deutsche Kurzfassung (German Abstract)

KI-Systeme werden zunehmend in Kontexten eingesetzt, in denen programmierte Entscheidungen ethische Implikationen haben und zentrale menschliche Werte betreffen. Beispielsweise müssen selbstfahrende Fahrzeuge entscheiden, wie viel Abstand sie zu anderen Verkehrsteilnehmenden halten oder, in extremen Fällen, mit wem sie in unvermeidbaren Unfällen kollidieren. Entscheidungen wie diese haben spürbare Auswirkungen auf Werte wie ‚Sicherheit‘ und ‚Fairness‘ und bestimmen letztendlich, was eine faire Risikoverteilung zwischen Verkehrsteilnehmenden bedeutet. Unternehmen müssen solche Entscheidungen im Voraus programmieren, um sicherzustellen, dass KI-Systeme ethische Aspekte explizit in ihre Entscheidungsfindung einbeziehen können, und schaffen somit künstliche moralische Agenten. Wie abstrakte Werte effektiv in KI-Systeme operationalisiert und manifestiert werden können, bleibt allerdings eine Herausforderung im Bereich der Maschinenethik. Daher untersucht diese Doktorarbeit, wie ethische Entscheidungsprinzipien in KI-Systeme (insbesondere selbstfahrende Fahrzeuge) integriert werden und welche gesellschaftlichen Auswirkungen daraus resultieren können.

Im Hinblick auf dieses Forschungsvorhaben besteht diese Doktorarbeit aus drei in sich geschlossenen Aufsätzen. Aufsatz I stützt sich auf qualitative Interviews mit Expert und Expertinnen aus der Philosophie, KI und Kognitionswissenschaften, um positive und negative Gründe für Computational Ethics und künstliche moralische Agenten zu untersuchen. Zusätzlich werden genannte Empfehlungen für den technologischen Entwicklungsprozess eines Unternehmens sowie Governance-Maßnahmen für die Industrie aufgeführt. Aufsatz II basiert auf einer systematischen Literaturrecherche im Bereich der Ethik des autonomen Fahrens und erörtert die Anwendbarkeit verschiedener ethischer Theorien auf die Entscheidungsfindung von selbstfahrenden Fahrzeugen. Aufsatz III stellt Systemanforderungen und ein konkretes Modell für die ethische Entscheidungsfindung selbstfahrender Fahrzeuge in gefährlichen Verkehrssituationen auf. Insgesamt zielt diese Doktorarbeit darauf ab, die Grundlagen für eine angewandte KI-Ethik zu legen, indem sie am Beispiel des selbstfahrenden Fahrzeuges zeigt, wie ‚ethische‘ künstliche moralische Agenten entwickelt werden können.

Table of Contents

Acknowledgments	I
Abstract.....	III
Deutsche Kurzfassung (German Abstract)	IV
Table of Contents	V
List of Figures.....	VIII
List of Tables	X
List of Abbreviations	XI
1 Introduction	1
1.1 Motivation and purpose of this dissertation	1
1.2 Theoretical background.....	4
1.2.1 Terms and definitions.....	4
1.2.2 AI systems and artificial moral agents	5
1.2.3 The case of self-driving vehicles.....	9
1.2.4 Machine ethics and integrating ethical principles	12
1.2.5 Related research fields and approaches: Subsumption & delimitation	16
1.3 Summary of research gaps and research questions	21
1.4 Research methods	22
1.5 Dissertation structure and summary of the three essays	26
References	28
2 Essay I: Formalizing Ethical Principles Within AI Systems: Experts' Opinions on Why (Not) And How to do it.....	40
2.1 Introduction	42
2.2 Theoretical background.....	43
2.2.1 (Designing) ethical decision-making of AI systems	43
2.2.2 Computational ethics	44
2.3 Methods.....	45
2.3.1 Participants.....	45
2.3.2 Data collection	46
2.3.3 Data analysis	47
2.3.4 Data validity.....	48

2.4	Findings.....	49
2.4.1	Artificial moral agents for (supporting) ethical decisions?	49
2.4.2	Computational ethics for developing artificial moral agents?.....	57
2.4.3	Recommendations for implementing computational ethics	70
2.5	Discussion: Facilitating computational ethics and artificial moral agents	84
2.6	Conclusion	87
	References	88
	Appendix A of Essay I – Interview guide.....	92
	Appendix B of Essay I – Frequency analyses of the experts’ arguments & recommendations	96
3	 Essay II: Applying Ethical Theories to the Decision-Making of Self-Driving Vehicles: A Systematic Review and Integration of the Literature	105
3.1	Introduction.....	107
3.2	Review method	110
3.2.1	Stage I: Identifying relevant literature	110
3.2.2	Stage II: Structural analysis of the literature	111
3.2.3	Stage III: Theme identification and article integration	111
3.3	Findings.....	112
3.3.1	Guiding ethical theories for the decision-making of SDVs.....	113
3.3.2	Advantages of applying particular ethical theories to the decision-making of SDVs.....	118
3.3.3	Disadvantages of applying particular ethical theories to the decision-making of SDVs....	126
3.3.4	Combined theories, additional considerations/principles & elaborate approaches for the decision-making of SDVs	135
3.4	Discussion	140
3.4.1	Key takeaways and implications from past literature	140
3.4.2	Critical remarks and updated research agenda.....	144
3.5	Conclusion	147
	References	148
	Appendix A of Essay II – Literature search & analysis.....	159
	Appendix B of Essay II – Descriptive/structural analysis of the literature	161
4	 Essay III: Ethical Decision-Making for Self-Driving Vehicles: A Proposed Model & List of Value-Laden Terms That Warrant (Technical) Specification ...	166
4.1	Introduction.....	168
4.2	Theoretical fundamentals	169
4.2.1	Risk (i.e., safety) assessment and management	170
4.2.2	Adjustments to the underlying calculation depending on the situation's criticality	171
4.2.3	Responsibility considerations and the protection of vulnerable road users.....	172

4.2.4	Implementing a mix of ethically grounded and socially shared principles	173
4.3	A proposed model for ethical decision-making of SDVs	175
4.4	Discussion	185
4.4.1	Benefits of this ethical decision-making model	185
4.4.2	Limitations, research agenda, and terms to be determined	187
4.5	Conclusion	189
	References	192
	Appendix of Essay III – List of relevant standards & regulatory documents	199
5 	Discussion	201
5.1	Summary of findings.....	201
5.2	Implications for society.....	203
5.3	Implications for practitioners	209
5.4	Implications for scholars.....	219
5.4.1	Contributions to research	219
5.4.2	Critical remarks, limitations, and future research agendas.....	221
	References	226
6 	Conclusion	230
	Appendix.....	231
	Appendix A: Reference & copyright information by the publisher for the first essay (Essay I, Chapter 2)	231
	Appendix B: Reference & copyright information by the publisher for the second essay (Essay II, Chapter 3)	232
	Appendix C: Reference & copyright information by the publisher for the third essay (Essay III, Chapter 4)	233
	Appendix D: Author contributions to the three essays in this dissertation	234

List of Figures

Figure 1: Level of automation outlined in SAE J3016 standard, adapted from SAE International (2021)	10
Figure 2: Self-driving vehicle architecture and decision-making structure, adapted from Pendleton et al. (2017) and Srivastava (2019)	11
Figure 3: Three basic layers of a values hierarchy, adapted from van de Poel (2013)	14
Figure 4: ‘Domains of ethics’ addressed in this dissertation, adapted from Sætra and Danaher (2022)	19
Figure 5: Research process and methods, adapted from Liu and Stephens (2019)	23
Figure 6: Experts’ indicated arguments for and against developing/utilizing artificial moral agents for ethical decision making (ordered by perspective and frequency)	56
Figure 7: Experts’ indicated arguments for and against utilizing computational ethics to develop artificial moral agents (ordered by perspective and frequency)	69
Figure 8: Experts’ indicated recommendations of how to implement computational ethics and artificial moral agents (ordered by type of recommendation and frequency)	83
Figure 9: Proposed model for facilitating computational ethics and artificial moral agents.....	86
Figure 10: Experts’ indicated arguments from a practical, societal, and epistemic view for developing/utilizing AMAs for ethical decision making (ordered by frequency)	96
Figure 11: Experts’ indicated arguments from a practical, societal, and epistemic view against developing/utilizing AMAs for ethical decision making (ordered by frequency)	97
Figure 12: Experts’ indicated arguments from a practical, societal, and epistemic view for utilizing computational ethics to develop AMAs (ordered by frequency)	99
Figure 13: Experts’ indicated arguments from a practical, societal, and epistemic view against utilizing computational ethics to develop AMAs (ordered by frequency)	100
Figure 14: Experts’ indicated recommendations for the company’s design and development of AMAs (ordered by frequency)	102
Figure 15: Experts’ indicated recommendations for the industry’s governance of AMAs (ordered by frequency)	103
Figure 16: Experts’ indicated recommendations for the scientific community investigating computational ethics and AMAs (ordered by frequency)	104
Figure 17: Data structure, reproduced from Corley and Gioia (2004)	112
Figure 18: Summarizing model of applying and integrating ethical theories to SDVs’ decision making.....	143
Figure 19: Search funnel of the systematic literature review, adapted from Moher et al. (2009)	160

Figure 20: Automated driving ethics publications over time and by discipline (2014-July 2023)	161
Figure 21: Utilized methodology of the identified publications	163
Figure 22: Practicability of the identified publications	164
Figure 23: Theory integration classified by practicability of the identified publications	165
Figure 24: Simplified traffic scenario	176
Figure 25: Proposed model summarizing the decision-making steps, guiding theories, underlying calculations, and terms that warrant (technical) specification	191
Figure 26: Overview of societal implications of integrating ethical principles into AI systems	208

List of Tables

Table 1: Summary of the essays.....	27
Table 2: Overview of interviewed experts	46
Table 3: Identified ethical theories in past literature that can guide SDVs' ethical decision making.....	114
Table 4: Identified advantages of applying ethical theories to the decision making of SDVs in past literature	125
Table 5: Identified disadvantages of applying particular ethical theories to the decision making of SDVs in past literature.....	134
Table 6: Overview of suggested combined theories, additional considerations/principles, and elaborate decision processes in past literature.....	139
Table 7: Overview of the literature search process	159
Table 8: Overview of research outlets by discipline (indicated by the assigned Web of Science category and/or Scopus subject area for the respective publication) and by journal/publication type.....	162
Table 9: Overview of risks (r_{Ai}, T_i) and an unspecified utility/objective (x_{Ai}, T_i) for all trajectory alternatives and the individual traffic participants.....	176
Table 10: Exemplary SDV duties distinguishing between non-hazard and hazard modes, adapted from Evans et al. (2020)	177
Table 11: Exemplary threshold restrictions for collision probability (c_{max}) and estimated harm (h_{max}).....	178
Table 12: Exemplary hierarchy and corresponding valence factors (v_i) for different types of traffic participants	179
Table 13: Overview of calculated valence-adjusted risks (vr_{Ai}, T_i) for all remaining trajectory alternatives and the individual traffic participants	179
Table 14: Exemplary weighting factors for risk inequality (w_E) and aggregated risk (w_U)....	180
Table 15: Overview of calculated risk inequality (E_{Ai}) and aggregated risk (U_{Ai}) for all remaining trajectory alternatives	180
Table 16: Overview of calculated principle-weighted risks (wr_{Ai}) for all remaining trajectory alternatives	181
Table 17: Definition and technical specification of key terms for the SDVs' decision steps ...	184
Table 18: List of standards and regulatory documents in the field of autonomous driving ethics that were consulted in Essay III	200
Table 19: Overview of the practical outputs and key takeaways of the three essays.....	218

List of Abbreviations

AI	Artificial intelligence
AMA(s)	Artificial moral agent(s)
AV(s)	Autonomous vehicle(s)
BAAI	Beijing Academy of Artificial Intelligence
BMJ	Bundesministerium der Justiz
BMW	Bayerische Motoren Werke
CAV(s)	Connected and automated vehicle(s)
DMV	California Department of Motor Vehicles
Ed.	Editor
Eds.	Editors
e.g.	exempli gratia
et al.	et alii
ETSI	European Telecommunications Standards Institute
EU	European Union
Gov.	Government
HLEG	High-level Expert Group on Artificial Intelligence
i.e.	id est
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
ISO	International Organization for Standardization
L3+	Driving automation of SAE Level 3 and more
MAIS	Maximum Abbreviated Injury Score
MAIS3+	MAIS score of 3 or more

n.a.	not available, not applicable
NACTO	National Association of City Transportation Officials
NHTSA	National Highway Traffic Safety Administration
NTC	National Transport Commission
OECD	Organisation for Economic Co-operation and Development
p.	page
PCSI	Potential crash severity index
pp.	pages
RQ	Research question
SAE	Society of Automotive Engineers
SDV(s)	Self-driving vehicle(s)
TUM	Technical University of Munich
TÜV	Technischer Überwachungsverein
UK	United Kingdom
UNECE	United Nations Economic Commission for Europe
U.S.	United States
V2V	Vehicle-to-vehicle
vs.	versus
VSD	Value-sensitive design
WHO	World Health Organization

1 | Introduction

1.1 Motivation and purpose of this dissertation

Looking at technological innovations such as ChatGPT, self-driving vehicles, or autonomous weapons, it is evident artificial intelligence (AI) systems are becoming increasingly sophisticated, autonomous, and adopted throughout various sectors and life-spheres (Dyoub et al., 2020). Next to routine tasks, these systems assist with or completely take over difficult decisions, some of which are ethically charged (Martinho et al., 2021) in the sense that they touch upon key human values and goals. One drastic example is self-driving vehicles (SDVs), which also represent a central topic of investigation in this dissertation. In most extreme cases, SDVs will have to decide over life and death by “choos[ing] what or who to crash into when an accident cannot be avoided” (Bonnemais et al., 2018; p.42). Decisions like these profoundly affect the values of ‘safety’ and ‘fairness’, essentially defining what constitutes a fair distribution of risk among traffic participants. As AI is not smart enough to naturally understand human suffering or to care about human goals or values (Coeckelbergh, 2020), ethical considerations have to be computed by their designers within the systems’ algorithms (Malle et al., 2015), and ideally, without implicitly reinforcing the designers’ own biases (Gogoll et al., 2021). This moves engineering and the design of AI systems away from a “dispassionate and objective activity” (Pirtle et al., 2021; p.10) and emphasizes the need for programmers to assume their role as active ‘value-smiths’ (Moor, 1995) or ‘practical ethicists’ (Verbeek, 2011). In fact, interviews with practitioners in related fields show that they already engage in explicitly incorporating ethical considerations when designing AI systems (Jacobs & IJsselsteijn, 2021).

These developments have elevated the subject of AI ethics “to the top of policy agendas for governments and other stakeholder groups at both national and international levels” (Mezgár & Váncza, 2022; p.394). For example, corresponding efforts include the key requirements for trustworthy AI developed by the High-level Expert Group on AI (2019), AI4People’s ethical framework for a good AI society (Floridi et al., 2021), OECD’s (2024) principles on AI, or IEEE’s (2019) 7000TM series of standards to address ethical considerations during system design. Also, many countries and governments have released sets of ethical guidelines, including the Blueprint for an AI Bill of Rights in the U.S. (The White House Washington, 2022), the Montréal Declaration for a Responsible Development of AI (2018) or Beijing AI Principles (BAAI, 2019). Similarly, companies such as BMW, Google, Microsoft, or SAP have published their own codes

of ethics for AI (e.g., BMW Group, 2020; Microsoft, 2022). Analyzing these various codes of conduct, guidelines, and principles reveals a convergence around themes and values such as fairness, non-discrimination, or safety (Fjeld et al., 2020). While these lists of high-level guidelines indicate ‘what’ to consider when designing AI systems responsibly, they only do so abstractly (Mittelstadt, 2019). For example, in terms of SDVs’ decision-making, it is stated the vehicles should demonstrate fair, unbiased internal functioning (Lütge et al., 2021) or that they should manage dilemmas by means of ‘shared ethical principles’ (European Commission, 2020), leaving open what this concretely entails. Due to this vagueness and abstraction, scholars have argued the existing corpus of AI ethics principles is relatively ‘meaningless’ (Munn, 2022) and so far has had a minimal effect on AI developing practices (McNamara et al., 2018). Instead, the “generic nature [of codes of conduct] leaves the reader with the feeling that their gut feeling and practical constraints should have the final verdict” (Gogoll et al., 2021; p.1097) and opens up the potential for companies to engage in ‘ethics washing’ (Munn, 2022; Shneiderman, 2020).

However, genuinely participating in the responsible development of AI can yield numerous benefits for companies. For one, it positively shapes consumer trust, satisfaction, and adoption by easing people’s worries about AI systems (Morley et al., 2021) as a result of them knowing the system’s behavior is constrained and guided by ethical principles (LaCroix & Luccioni, 2022). This will also help justify the AI system’s reasoning and behavior once a detrimental outcome is caused (Nyholm, 2023c). Therefore, integrating ethical principles into AI systems should be in the self-interest of companies seeking to ensure the adoption of their systems (Lütge, 2024) or to proactively address forthcoming liability issues. Alongside these self-motivated incentives, companies may soon be obligated to pursue an ethically aligned design of AI systems and disclose their respective efforts. For example, companies will have to carry out risk assessments, establish appropriate mitigating strategies (e.g., ISO, 2022; European Commission, 2021), and document their corresponding efforts in repositories such as the ‘Value Register’ (e.g., IEEE, 2021). Thus, it is a timely and pressing matter to bridge the gap between AI principles and AI practice.

Turning ethical guidelines into practice demands their instantiation within AI systems, for example, through operationalizing high-level values into concrete and tangible parameters, system functionalities, or design requirements (Brey & Dainow, 2021; Segun, 2021; Spiekermann, 2023; van de Poel, 2013). Regarding SDVs – one example of AI systems that make ethical decisions – these design requirements extend to the systems’ underlying logic and reasoning process. A system’s capacity for ethical reasoning can be facilitated through “algorithms or other features that enable them to explicitly make use of ethical considerations in their decision making” (Nyholm, 2023c; p.163). By doing so, so-called ‘artificial moral agents’ (AMAs) are created, which hopefully ‘do the right thing’ based on the ‘right motivations’ that are

sensitive to ethical considerations and align with human values (Nyholm, 2023c). In this context, doing the ‘ethical’ thing or doing ‘good’ would mean “supporting the realization of positive value [such as increased safety in traffic] or the reduction of negative values [such as discrimination among road users]” (Spiekermann, 2023; p.83; IEEE, 2021). To achieve this, machine ethicists have yet to determine how precisely to structure the process of integrating ethical principles into AI systems’ reasoning process (Woodgate & Ajmeri, 2022) and how to ensure that AI systems respect the embedded principles (van de Poel, 2020). Additionally, the context of SDVs poses the questions of what constitutes a ‘fair’ decision (Awad & Levine, 2020) and what exactly are these ‘shared ethical principles’ that shall guide SDVs’ decision-making (compare European Commission, 2020).

In this endeavor, it is key to pay attention to the societal ramifications that result from pertinent AI systems, especially since their decisions/actions will have a greater sphere of influence and will be more systematic compared to when human beings make ethical decisions individually (Song & Yeung, 2022). However, looking at the rapid adoption of complex AI systems, “there seems to be a real risk that in the near future [...] we will insufficiently understand their ethical and societal [negative] implications and nevertheless use them widely” (Coeckelbergh, 2020; pp.15-16). For example, SDVs are said to increase road safety, but due to software bugs or sensor malfunctioning, they can still lead to fatal crashes (Lütge et al., 2021). Similarly, less obvious secondary effects may emerge depending on how an AI system is designed. For instance, SDVs that aim to maximize social welfare with respect to survivability in traffic accidents may learn to strike motorcycle riders wearing helmets as opposed to those not wearing any. This represents implicit discriminatory targeting, which may, in turn, incentivize motorcycle riders not to take relevant safety precautions (Caballero et al., 2023). This example illustrates that the programming of SDVs is not straightforward and that “formal ethics are an instrument of power that can be exploited to perpetuate particular moral norms” (Hoffmann, 2021; p.323). Therefore, it is important to investigate and understand the societal implications of AMAs and their underlying logic during the development stage proactively rather than complain afterward about problems that arise as a result (Coeckelbergh, 2020). Specifically for ‘high-risk’ AI systems, the EU AI act prescribes a strict obligation to conduct rigorous risk assessments and to put mitigation measures in place before they can be put on the market (European Commission, 2021). Only the proactive identification of ramifications facilitates the timely development of countermeasures (e.g., in the form of a specific way of programming a system’s algorithm), if necessary (Wallach & Allen, 2009).

To address the sketched issues, the purpose of this doctoral dissertation is twofold. First, it aims to investigate the societal consequences of (not) integrating ethical principles into AI

systems, with a particular focus on SDVs (see Research question 2 on p.21). Second, it aims to engage in the practical investigation of how to do so (see Research question 1 on p.21). The ‘societal investigation’ goes hand in hand with the ‘practical investigation’. After all, recognizing and defining risks (including those related to societal ramifications) is essential before formulating ethical guidelines, devising suitable countermeasures, and establishing responsible AI (Mezgár & Váncza, 2022). Before addressing these research inquiries in Chapters 2 | through 5 |, the following sections will provide an overview of the theoretical background, the concrete research questions, the utilized research methods, and the structure of this dissertation.

1.2 Theoretical background

1.2.1 Terms and definitions

To establish a clear and coherent basis for understanding, this section defines key terms for this dissertation. These definitions are derived from past literature.

Artificial Intelligence (AI) systems: “software (and possibly also hardware) systems [...] that, given a complex goals, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal” (HLEG, 2019; p.6). AI systems can be ‘hardcoded’ by following symbolic rules and/or they can learn independently from data or experience (ISO, 2022).

Artificial moral agents (AMAs): “technologies that are equipped with algorithms or other features that enable them to explicitly make use of ethical considerations in their decision-making” (Nyholm, 2023c; p.163). This definition corresponds to Moor’s (2006) conception of ‘explicit ethical agents’.

Ethical² AI: “refers to the development, deployment and use of AI that ensures compliance with ethical norms, including fundamental rights as special moral entitlements, ethical principles and related core values” (Brey & Dainow, 2021; p.42).

Machine ethics: “the interdisciplinary project of trying to design and build ‘artificial moral agents’: technologies able to make moral decisions” (Nyholm, 2023a; p.5).

² In line with contemporary scholars (e.g., Cave et al., 2018; Sætra & Danaher, 2022; Singer, 2011), the terms ‘ethical’ and ‘moral’ or ‘ethics’ and ‘morality’ will be utilized interchangeably in this dissertation.

Ethics by Design: “the technical/algorithmic integration of ethical reasoning capabilities as part of the behavior of artificial autonomous systems” (Dignum, 2018; p.2). It constitutes one method in AI ethics to achieve ethical AI, focusing on the systems themselves rather than the humans who develop them (Dignum, 2018).

Computational Ethics: “seek[s] ways to translate abstract moral principles into computer codes [...] [and to] develop ethically grounded algorithms” (Segun, 2021; pp.270-271).

Ethical decision-making: “a process by which individuals use their moral base [and intuition] to determine whether a certain issue is right or wrong” (Carlson et al., 2009; p.536; Kahneman, 2003). When it comes to AI systems, as opposed to human beings, ethical decision-making entails the process by which the systems use ethical principles to calculate and choose whether a particular course of action is right or wrong (Anderson, 2011).

Ethical (decision-making) principles: “operationalizable rules inferred from philosophical theories [and AI principles/guidelines], which guide decision-makers in making normative judgments and determining the moral permissibility of concrete courses of actions” (Woodgate & Ajmeri, 2022; p.3).

Ethical framework³: “[...] gives us a way for dealing with situations involving ethical dilemmas thanks to principles, metrics, etc.” (Bonnemains et al., 2018; p.46).

Self-driving vehicles (SDVs): a motor vehicle that operates within a predefined operational design domain without the human in the car intervening or performing any driving tasks (at least for a certain period). This refers to SAE levels at or beyond Level 3 (SAE International, 2021).

Trajectory planning: “is a significant component of autonomous vehicle systems, which directly influences the automated traffic safety” (Wang et al., 2019; p.8546). Trajectory planning involves computing a safe path from the SDV’s present position to a desired target state, considering the surroundings and projected movements of other traffic participants (Wei et al., 2015).

1.2.2 AI systems and artificial moral agents

A brief introduction to artificial intelligence. Artificial intelligence (AI) systems are software or hardware systems that act in physical or digital environments, interpret and process (collected) information to reason, and decide the best action(s) to achieve specific goals (HLEG, 2019; Haenlein & Kaplan, 2019). The basis of the ‘intelligence’ of an AI is its underlying algorithm, which “is a set and sequence of instructions, like a recipe, that tells the computer, [...] machine,

³ In this doctoral dissertation, ‘framework’ and ‘model’ will be used interchangeably.

robot, or whatever it is embedded in, what to do” (Coeckelbergh, 2020; p.70). The path that AI systems follow either can be ‘hardcoded’ or shaped through machine learning techniques (ISO, 2022). AI systems that are hardcoded resemble expert systems that rely on the explicit design of a decision procedure (e.g., in the form of if-else statements) (Evans, 2021). When it comes to machine learning, the steps needed to solve a certain problem are not determined a priori, but AI systems learn independently from data or experience (ISO, 2022). Different types of machine learning methods exist, such as supervised, unsupervised, and reinforcement learning. Commonly seen in machine learning methodologies, AI systems build statistical models based on large datasets (images, driving behavior, etc.) to identify recurring patterns and correlations, make predictions in new situations, and subsequently select the best course of action. This process can be supervised, utilizing labeled datasets that pair input data with the correct output data, or unsupervised, where the AI system independently uncovers patterns in unlabeled data (Evans, 2021). Reinforcement learning methods “involve goal-directed learning from interaction with an environment guided by the use of a reward signal that the machine attempts to maximize” (Kaas, 2021; p.58). Through an iterative process of trial and error, those actions that yield a positive reward are reinforced, while those that yield a negative reward are omitted by the AI system (Evans, 2021). While algorithms predominately used to function based on hardcoded decision trees (‘top-down’ approaches), nowadays, the dominant paradigm in AI systems is rather machine learning (‘bottom-up’ approaches) (Evans et al., 2023) or a combination of both (‘hybrid’ approaches) (Coeckelbergh, 2020). Similarly, the cognitive architecture of most embodied artificial agents is hybrid. For example, SDVs might utilize bottom-up approaches for perception tasks, such as recognizing and classifying their environment while employing top-down approaches for tactical (potentially ethical) decision-making (Evans, 2021).

Ethical decision-making of human beings and AI systems. As AI systems become more sophisticated and ‘autonomous’⁴, they are increasingly “able to operate on their own, performing specific tasks, without direct human steering” (Nyholm, 2023c; p.14), so the tasks they can perform even include making complex and ethical decisions (Martinho et al., 2021). For instance, in healthcare, algorithms decide how scarce medical resources such as ventilators or donated organs are allocated. In the criminal justice system, algorithms determine the sentences for perpetrators. Self-driving vehicles decide how to distribute risks between traffic participants (Awad & Levine, 2020). These examples show that there are ethical ramifications to what AI

⁴ In general, this dissertation will refrain from using the word ‘autonomous’ as much as possible, but if utilized, the term here is not understood in a strict philosophical sense, whereby complete freedom of action is required (Hunyadi, 2019). Instead, it describes systems that “when in a given domain and for given tasks, they are capable of accomplishing the specified tasks on their own despite some changes in this domain’s environment” (Evans et al., 2023; p.3).

systems do as they “can help or harm, or produce good or bad outcomes” (Nyholm, 2023c; p.161) and “undoubtedly feed back into, and thus change, the existing moral ecology” (Wallach & Allen, 2009; p.62). Namely, referring to the previous examples, AI systems can help or harm the health situation of patients, affect the personal freedom of perpetrators or the safety/criminality levels in society, as well as shape the physical integrity of traffic participants. However, ethical behavior is not only about the manifestation of (positive) tangible outcomes but also about the underlying moral process leading up to the decision to act in a certain way (Hunyadi, 2019).

Human ethical decision-making is defined as “a process by which individuals use their moral base [and intuition] to determine whether a certain issue is right or wrong” (Carlson et al., 2009; p.536; Kahneman, 2003). Rest’s (1986) four-component model represents a traditional tool to describe human deliberate ethical decision-making and behavior. The model comprises the following steps: moral awareness, moral reasoning, moral intent, and moral behavior. First, individuals need to identify an ethical issue (moral awareness) (Rest, 1986), which, amongst others, is influenced by the moral intensity of a situation (Jones, 1991). Subsequently, individuals engage in moral reasoning, extracting, weighing, and integrating morally relevant information and considering various standards and ethical principles (Bandura, 1991). Following, individuals develop a moral intent by prioritizing certain moral concerns ahead of others (Rest, 1986; Campbell, 2017). As a last step, individuals perform moral behavior, meaning they take actions aligned with their moral intent (Rest, 1986). Ethical decision-making of AI systems constitutes the process by which the systems analyze and choose whether a certain issue or course of action is right or wrong. Analogous to the deliberate ethical decision-making of humans, SDVs would need the capacity to recognize an ethical issue in the first place and subsequently address the issue based on consulted ethical principles (and/or experience⁵). Thus, putting AI systems in such contexts where they engage in actions that cause good or evil (i.e., morally qualifiable actions) (Verbeek, 2011), requires them to be sensitive to and account for ethical considerations by satisfying appropriate norms in their ethical decision-making process (Etienne, 2022). Another central requirement for decision-making, especially in algorithmic decision-making, is consistency, which implies that “cases that share the same relevant characteristics should result in similar decisions” (Awad et al., 2022; p.393). According to Kim et al. (2020), “[e]thical principles are necessary conditions for the coherence [...] of the reasons behind an action” (p.10). This necessity likely extends to the coherence of the subsequent decisions made, which

⁵ Extending the rationalist theories of ethical decision-making, contemporary literature acknowledges the predominance of intuition, affects, or heuristics when it comes to human ethical decision-making (Haidt, 2001; Kahneman, 2003; Zollo et al., 2017). Although capabilities such as affects or intuitions are typically associated with human beings, AI systems have also been considered to possess a form of ‘artificial intuition’ due to employing machine learning techniques (Trovati et al., 2022).

underscores the significance of systematically programming AI systems' reasoning process (including the consulted ethical principles).

AI systems as artificial moral agents. As illustrated by the earlier examples, AI systems can be considered some kind of 'moral agents' that are capable of making ethical decisions and acting on those decisions (Moor, 2006; Nyholm, 2023c). "A good moral agent is one that can detect the possibility of harm or neglect of duty, and can take steps to avoid or minimize such undesirable outcomes" (Wallach & Allen, 2009; p.16). Technically, this can be achieved by equipping the systems "with algorithms or other features that enable them to explicitly make use of ethical considerations in their decision-making" by, for example, following a set of specific rules and ethical principles (Nyholm, 2023c; p.163) (see more in 1.2.4). Such an underlying decision procedure makes up an AI system's 'artificial morality' (Evans, 2021). The ultimate aim is that these systems will then produce the right answers (i.e., an output) based on the input and using the rules provided, but without possessing a genuine understanding or cognitive states (Coeckelbergh, 2020). There has been an ongoing debate about the extent to which artificial agents, such as AI systems, can be genuinely moral or moral agents (like human beings are) since they lack consciousness, intentionality, free will, or the ability to bear moral responsibility (van de Poel, 2020). Furthermore, it has been argued that prevalent efforts in machine ethics transport an oversimplified notion of ethical decision-making and behavior, as ethics cannot be entirely and easily codified into a set of simple and clear rules (Schwarz, 2023). However, "[i]t would be shortsighted and dangerous to dismiss the problem of how to design morally sensitive systems on the grounds that it is not genuine moral agency" (Wallach & Allen, 2009; p.199). Essentially, AI systems will carry out actions that have ethical impacts, effectively functioning as 'implicit ethical agents' (Moor, 2006), irrespective of whether they are deliberately programmed for it or possess a genuine understanding of (the reasons behind the) ethical decision-making (Evans et al., 2023).

Therefore, this doctoral dissertation will not focus on answering whether AI systems can be 'genuine' or 'full' moral agents and does not aim to convey an anthropomorphic vision of AI (decision-making). Instead, in line with contemporary scholars in the field of machine ethics, ethical decision-making or reasoning is here understood in a less demanding and "broader sense to mean simply whatever processing is carried out to reach a[n ethical] conclusion" (Cave et al., 2018; p.564). When referring to AMAs, this dissertation will assume the notion of so-called 'explicit ethical agents', meaning these are systems that will be (required to) act in "ethically salient contexts" (Evans, 2021; p.39) and are "programmed to resolve 'ethical' problems in line with explicit moral principles" (Hunyadi, 2019; p.2). Thus, this doctoral dissertation aims to show how to "build in' specific forms of morality" (Verbeek, 2011; p.2) and, thereby, design ethically sensitive systems that engage in some kind of ethical reasoning (Bonnemais et al., 2018) and align

with certain human values (Fossa, 2023). To effectively study the feasibility and ethicality of AMAs, it is useful to concentrate on one specific technological application (Danks, 2022), such as self-driving vehicles.

1.2.3 The case of self-driving vehicles

Levels of driving automation. As most road accidents are caused by human errors such as negligence or drunk driving, self-driving vehicles (SDVs) are expected to substantially improve road safety (Fossa & Cheli, 2023), pushing forward the advancement of driving automation. The driving automation of cars has been clustered into six levels spanning from Level 0 (i.e., no driving automation) to Level 5 (i.e., full driving automation) (illustrated in Figure 1) (SAE International, 2021). While in Level 0, the human driver performs all driving tasks, in Level 5, the vehicle performs all driving tasks under all road conditions, without the need for a human driver to take over at any point (SAE International, 2021). The levels in-between (1-4)⁶, “require some [and a varying] degree of on-board supervision[, for example,] in challenging conditions such as storms, heavy urban traffic, or imminent collision” (Evans, 2021; p.314). In this dissertation, the term “‘self-driving vehicles’ will here be used to refer to automated cars that can operate in so-called autonomous mode either in all traffic scenarios or in at least some traffic scenarios” (Nyholm, 2023c; p.53). The ‘autonomous mode’ denotes that the SDV can perform certain tasks and drive (for a period at least) without requiring any intervention or involvement in the driving tasks from the human inside the car (Nyholm, 2023c). This mode applies to SDVs at or above the level of 3. Here, the drivers are not actively engaged in (m)any driving tasks; instead, in most traffic situations, the vehicle independently performs the entire dynamic driving task, which potentially includes making (moral) decisions (SAE International, 2021). Therefore, this dissertation will focus on SDVs at or above the automation Level 3.

⁶ In detail, the six levels of automation are as follows: Level-0 vehicles constitute conventional cars where the human driver performs all driving tasks manually. At this level, the vehicle provides minor support to the driver with features such as lane departure warnings. Level 1 represents the lowest level of automation, offering only a single assistive feature such as lane centering or adaptive cruise control. At Level 2, multiple assistive features can run simultaneously; for example, the vehicle can provide steering, braking, and acceleration support to the driver. Vehicles at this level are not classified as ‘self-driving’ since the human driver is required to constantly supervise the support features and take over control at any point. At Level 3, the vehicles are equipped with genuine ‘automated driving features’ (as opposed to ‘driver support features’). These features allow the vehicle to drive without human involvement in particular conditions. Human drivers can take their eyes off the traffic situation but need to be able to take over control if requested by the vehicle. For example, Level-3 vehicles could serve as traffic jam chauffeurs on the highway, but the human driver would still need to drive the vehicle to and from the highway. Level-4 vehicles are highly automated and capable of operating without human intervention within predetermined operational design domains (defined by a set of environmental or geographical conditions etc.). Examples of these vehicles include driverless taxis that travel inside designated areas, such as urban environments with restricted speed limits or under favorable weather conditions. Level-5 vehicles perform all driving tasks in all conditions and without geofencing limitations (SAE International, 2021).

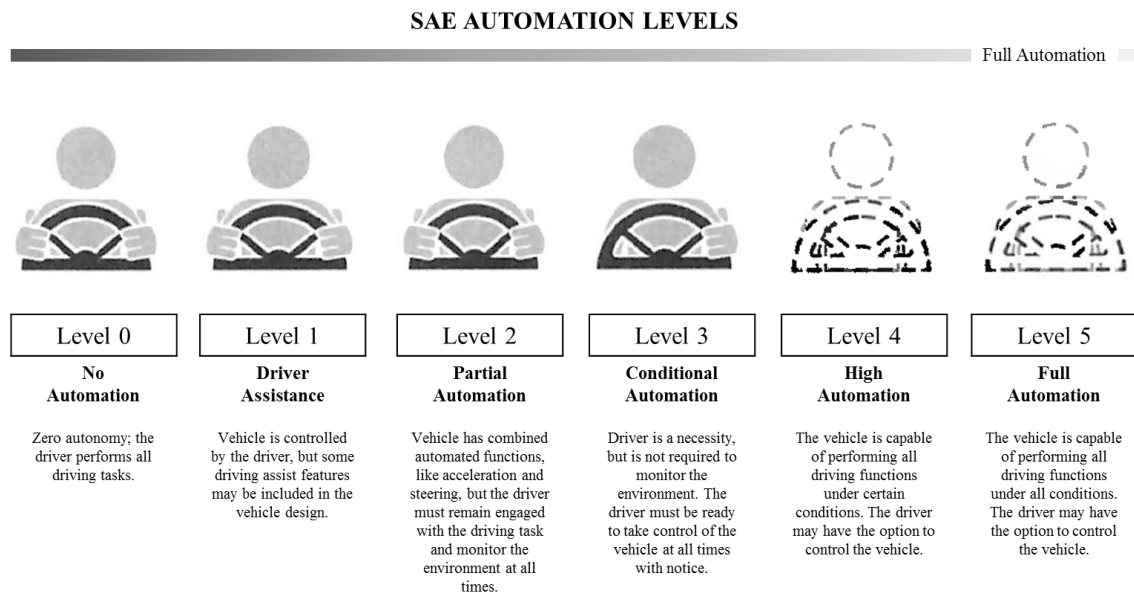


Figure 1: Level of automation outlined in SAE J3016 standard, adapted from SAE International (2021)

A brief overview of self-driving car developments. SDVs of such advanced levels (at least Level-3 cars) are already a reality. Over the past decades, driving automation has steadily increased, turning scientific ideas and theories of driverless cars into actual prototypes that are tested and launched. For example, in 2009, Google initiated ‘Waymo’ as a driverless car project (which later became its own technology company) in which employees consistently worked on the development of self-driving software and hardware (Deemantha & Heetige, 2022). In 2015, Tesla introduced its ‘Autopilot’, an advanced driver-assistance system that allows automatic lane changing, auto steering, or self-parking (Hull, 2015), which falls into the category of automation Level 2. Since 2017, driverless shuttles and taxis have been commercialized by companies such as Waymo or Uber and drive – with or without safety drivers – on roads, for example, in Arizona or California (Ahmed et al., 2022). In 2021, Honda became the world’s first automaker to sell a production vehicle of Level-3 in Japan that does not require constant monitoring by the driver (Beresford, 2021). More recently, it was announced that Mercedes-Benz, with its ‘Drive Pilot’ system, was authorized to introduce Level-3 standard-production vehicles on public freeways in the U.S. (Edward, 2023). So far, although no company, including Waymo or Tesla, has commercially produced a Level-5 car, they indicated their interest and aspiration to do so in the future (Kosuru & Venkitaraman, 2023). Overall, it is predicted that by 2035 up to 57% of new passenger cars sold may be self-driving cars with L3+ autonomous driving functions (McKinsey & Company, 2023). Therefore, the phenomenon of SDVs taking over decisions in road traffic is already prevalent today and will only increase over time.

Decision-making of self-driving cars. To be able to make decisions, SDVs need an underlying software architecture that allows them to sense, plan, and act (Coskun, 2021; Srivastava, 2019) (illustrated in Figure 2). First, to be able to operate in an environment, robots such as SDVs need the ability to sense their surroundings in real-time, such as the presence of obstacles or free space detection (Srivastava, 2019) through input data from cameras, LiDAR, or other sensors (Wei et al., 2015). Once the sensed information is collected and the SDV's position relative to its surroundings is determined, the SDV has to develop a strategy for action in the planning stage (Srivastava, 2019). For example, based on predictions of all detected objects and given existing vehicle/environmental constraints, within the trajectory planning of SDVs, potential trajectories are identified, and the 'best' trajectory is calculated (Wei et al., 2015). The motion control system translates this planned trajectory into action by commanding the SDV's actuators to cause a certain motion (Srivastava, 2019). For example, an SDV may slow down, accelerate, change its lane, or brake based on its calculation that this action will lead to collision avoidance (ETSI, 2019). Particularly in the process of trajectory planning, SDVs have to make difficult ethical decisions on their own, such as determining how to "balance the distribution of unavailable risk across everyone on the road – drivers, passengers, pedestrians, cyclists" (Awad & Levine, 2020). These decisions are made throughout the operation and at any point the vehicle moves on the street: in mundane traffic situations, in which the vehicle chooses how much space is given to individual traffic participants, as well as in critical accident situations, in which the vehicle chooses who to hit (Geisslinger et al., 2021). Thus, the programming of SDVs has continuous and direct effects on all traffic participants in terms of risk exposure or traffic victims (Mordue et al., 2020) and, thus, needs to be sensitive to moral reasons (Nyholm, 2023c). This is why SDVs can be clustered into the category of artificial moral agents (Segun, 2021), as defined earlier, which puts the programming of SDVs' decision-making under particular scrutiny.

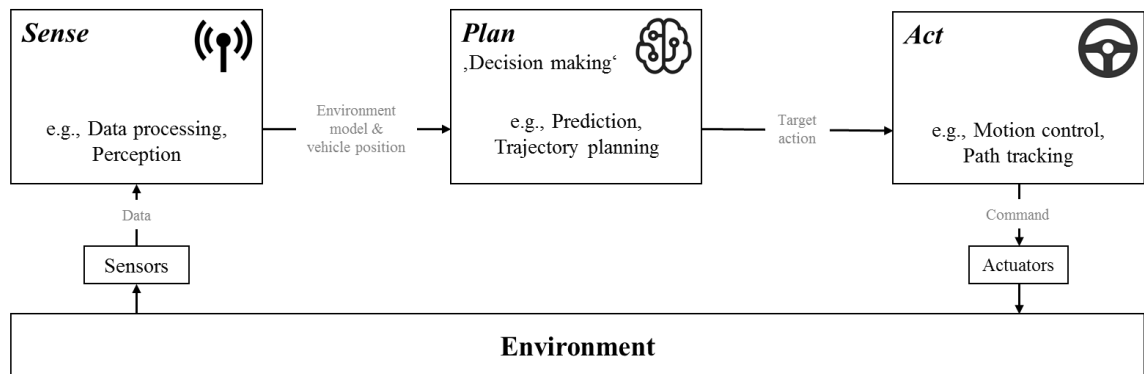


Figure 2: Self-driving vehicle architecture and decision-making structure, adapted from Pendleton et al. (2017) and Srivastava (2019)

Efforts to govern ethical decision-making of self-driving cars. In the past, a few automobile manufacturers or technology companies have put forward some controversial approaches to programming SDVs' decision-making. For example, in 2016, a representative from Mercedes-Benz announced their cars would always prioritize the safety of their passengers over external road users such as pedestrians (Taylor, 2016). In 2014, Google released a patent describing lateral lane positioning methods for SDVs to minimize their own risk exposure (Dolgov & Urmson, 2014). As such SDV programming focuses on the humans inside the vehicle, they “possibly [entail] devastating consequences for vulnerable road users” (Geisslinger et al., 2021; p.1043). To undermine such discriminatory programming, policymakers – akin to efforts targeting the governance of AI systems in general – have started to recognize the importance of acknowledging and governing the ethical aspects integral to SDV algorithms. Initiatives include, amongst others, the German Ethics Code for automated and connected driving (Lütge, 2017), the Ethics of connected and automated vehicles developed by a working group of the European Commission (2020), or the European Commission's (2022) legislation for the type-approval of the automated driving system of fully automated vehicles. In the United States, several bodies have encouraged developers and manufacturing companies to ensure SDVs obey traffic laws, follow ‘reasonable’ road etiquette (NHTSA, 2017), and improve traffic safety (National Conference of Commissioners on Uniform State Laws, 2019)⁷. Also, automobile manufacturers themselves (including those previously cited as ‘poor’ examples) have established own responsible AI principles and practices more recently (e.g., Mercedes-Benz Group, 2024). Although these efforts intend to counteract the ‘wild west’ of SDV development (Koopman & Widen, 2023), they still lack effectiveness at the current stage in the sense that it is not clear how these ethical and legal documents can be translated into formal rules for utilization by SDVs (Bin-Nun et al., 2022).

1.2.4 Machine ethics and integrating ethical principles

The project of machine ethics. Considering the technological applications outlined in previous sections, especially SDVs, it seems necessary and inevitable for some AI systems to engage in ethical decision-making. This underscores the urgency to explore methods for developing such moral machines. This, amongst others, is what investigations within the interdisciplinary research field called ‘machine ethics’ are devoted to, namely, designing and building technologies (i.e., artificial moral agents) that are able to process moral information, compute moral choices and correspondingly make moral decisions (Anderson & Anderson, 2011; Nath & Sahu, 2020; Nyholm, 2023a). The closely related approach of ‘Ethics by Design’

⁷ See a more comprehensive list of existing standards and regulations in Appendix of Essay III – List of relevant standards & regulatory documents.

addresses the ethics of the behavior of AI systems (Dignum, 2018) and emphasizes efforts to provide “the systems themselves with the capacity for explicit moral reasoning and decision-making” (Wallach & Allen, 2009; p.39). Similarly, computational ethics “seek[s] ways to translate abstract moral principles into computer codes [...] [and to] develop ethically grounded algorithms” (Segun, 2021; pp.270-271). Overall, these endeavors are concerned with developing suitable constraints on system behavior as well as integrating moral values and ethical reasoning capabilities within the algorithms (i.e., a machine’s ‘soul’) that ultimately determine the systems’ actions (Dignum, 2018; Evans et al., 2023). For core values to become integral elements of AI systems, they must be practically instantiated within the systems in the form of design/system requirements (Brey & Dainow, 2021).

From values to norms to design requirements. A key challenge in the field of machine ethics is creating systematic ways of encoding values into AI systems (Woodgate & Ajmeri, 2022), so that “[m]orality then, in a sense, becomes part of the intended ‘functionality’ of the product” (Verbeek, 2011; p.91). Thus, a crucial step involves converting values into tangible design requirements, as illustrated in Figure 3, which depicts the three basic layers of a values hierarchy (van de Poel, 2013). First, values constitute things humans consider important and meaningful in life (Friedman et al., 2013). They are generally more universal and stable than individual preferences/opinions (Abramson & Inglehar, 1995; van de Poel, 2009). Examples of values include human welfare, privacy, autonomy, and freedom from bias (Friedman et al., 2013). AI systems that violate such values may be considered ‘unethical’ (Brey & Dainow, 2021). Furthermore, values can be positive (e.g., fairness) or negative (e.g., discrimination). Technological applications, as so-called ‘value bearers’, have the potential to carry and, thereby, reinforce both positive and negative values (IEEE, 2021). For example, “in the transport domain, intelligent speed assist sets a maximum to the speed, strengthening the value of safety, but limiting the autonomy of the driver” (Flipse & Puylaert, 2018; p.51). The intermediate layer of the values hierarchy comprises ‘norms’, which encompass “all kinds of prescriptions for, and restrictions on, action”, thereby representing a specification and holding a deontic (or prescriptive) domain in comparison to the higher-level values (van de Poel, 2013; p.258). These norms are context-dependent. For example, if a technological application poses a risk of emitting toxic substances, the value ‘safety’ may be specified as ‘minimizing the amount of toxic releases from the system’ (van de Poel, 2013). By contrast, safety in the context of autonomous driving may be specified as ‘minimizing the number of traffic fatalities’.

Lastly, design requirements form the most concrete layer of a values hierarchy and extend to requirements for the systems itself and the process by which the system is developed (Sclove, 1995; van de Poel, 2013). For example, to reinforce the value of safety within/through SDVs, an

appropriate system requirement might entail an algorithmic feature that consistently ensures a safe distance from other objects and participants in traffic (compare Essay III in Chapter 4 |). Therefore, to manifest values within AI systems, a key design requirement pertains to the logic and decision-making process employed by the system. Especially when it comes to SDVs, it is essential that the system’s decision-making draws on “an ethical principle or set of principles [...] to calculate[, for example,] how it ought to behave in an ethical dilemma” (Anderson, 2011; p.22). These principles then provide comprehensive guidance for the systems’ decision-making (Woodgate & Ajmeri, 2022) so that they can assess alternative courses of action from different evaluative perspectives and with respect to certain moral criteria (Wallach & Allen, 2009).

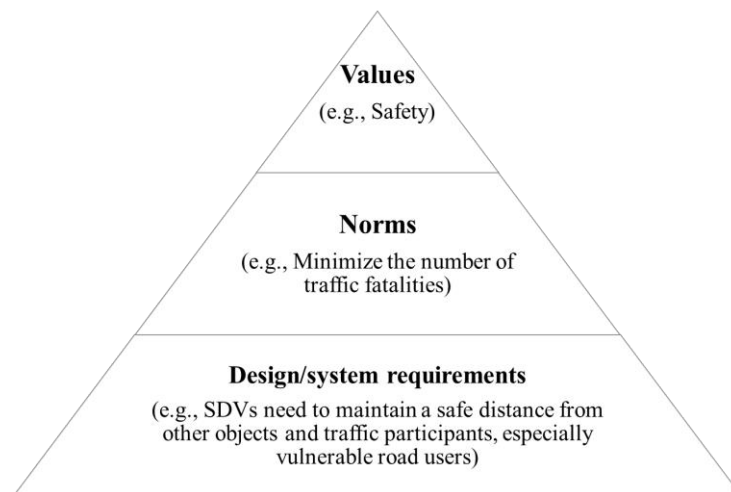


Figure 3: Three basic layers of a values hierarchy, adapted from van de Poel (2013)

Ethical decision-making principles & their (philosophical) underpinnings. A key challenge for programmers is to identify what constitutes ‘ethically correct’ (Bringsjord et al., 2006) and ‘fair’ decisions (Awad & Levine, 2020) so that these AI systems do not simply become an instrument of power to replicate a particular moral view (Hoffmann, 2021) or human biases (Pirtle et al., 2021). Thus, the endeavor of machine ethics raises the normative issue of deciding what values and related ethical principles should be implemented (Nyholm, 2023a). For one, it has been argued that ethical decision-making principles employed by AI systems have to align with deeply held human values (Nyholm, 2023a) and fundamental human rights such as the ones emphasized in the existing body of AI principles or guidelines (e.g., fairness, safety). In fact, previous studies have illustrated how values and norms can be derived or specified by drawing on the Ethics Guidelines for Trustworthy AI or the AI for Social Good principles (e.g., Umbrello & van de Poel, 2021).

Next to these new moral ideas (e.g., AI principles) or legal principles (Scheutz & Malle, 2017), academics contend that we may not need to reinvent the wheel completely. Namely, philosophy and general ethical theory can serve as a foundation or at least as one part of the solution in machine ethics (Nyholm, 2023c). After all, ethics is the “branch of knowledge or theory that investigates the correct reasons for thinking that this or that is right” (IEEE, 2021; p.18) and offers many theories and tools that have been studied for thousands of years and may be informative when designing AI systems (Cawthorne, 2022). In particular, falling back on these insights from philosophy (moral philosophy and ethics specifically) can “broaden our perspective on a debate, to look at it from a wider lens [...] [and] add a layer of rigorous principled thinking to value-laden discussions” (Bietti, 2020; p.214). Similarly, Gryz (2020) argues investigating and ultimately choosing an appropriate theory of normative ethics is one of the first decisions a designer of an AMA has to make. Nevertheless, in the current state, the field of AI ethics still lacks philosophical underpinnings (Bakiner, 2023). Some traditional ethical theories that were mentioned as candidates for computational use for AI systems include deontology, consequentialism, and virtue ethics (Kim et al., 2021), which all offer different lenses and implications that may need to be factored into the choice of AI systems (Wallach & Allen, 2009). In this dissertation, I will emphasize a philosophy-informed approach to machine ethics by assuming that ethical principles can be inferred from, amongst others, philosophical theories such as utilitarianism (Woodgate & Ajmeri, 2022). Ethical (decision-making) principles are, therefore, here understood as “operationalizable rules inferred from philosophical theories [and AI principles/guidelines], which guide decision-makers in making normative judgments and determining the moral permissibility of concrete courses of actions” (Woodgate & Ajmeri, 2022; p.3).

Technical programming approaches in machine ethics. In line with the previously indicated technical programming methods of AI (see 1.2.2), ethical principles can be technically integrated into AI systems in a top-down, bottom-up, or hybrid fashion (Etzioni & Etzioni, 2017). In the top-down approach, ethical principles or rules are preprogrammed into the machine’s guidance system (Gentzel, 2020), which seems exceptionally well suited for implementing machine ethics (Kaas, 2021). This means existing ethical theories are transformed into algorithms instructing the machine to behave/decide in a particular manner during operation (Lucifora et al., 2021). For example, an SDV’s decision-making could be programmed to function in a utilitarian fashion, such choosing to run over an older person instead of a child when confronted with action alternatives (Gentzel, 2020). Conversely, as mentioned previously, the bottom-up approach somewhat resembles intuitive decision-making. Namely, no formal rule or ethical theory is directly implemented. Instead, the systems are provided “with environments and instances to learn

preferable behaviors and infer core values to arrive at moral decisions and actions” (Zoshak & Dew, 2021; p.2). Technically, this learning mechanism is realized either through a neural network by observing and replicating actual human behavior or through reinforcement learning by rewarding machines for their moral behavior (e.g., Etzioni & Etzioni, 2017; Goodall, 2014; Lucifora et al., 2021; Roff, 2018; Siegel & Pappas, 2021). For example, SDVs “could be trained on a combination of simulation and recordings of crashes and near crashes, with human feedback on the ethical response” (Goodall, 2014; p.62). While the decisions taken through machine learning techniques cannot fully be determined or controlled ex-ante, the humans behind the technology (i.e., the developers) still frame the overarching rationale of such AI systems by selecting specific training data sets and variables or by defining the reward/cost function that stands to be optimized (Agrawal et al., 2018; TÜV Süd, 2023). However, because they are associated with issues like perpetuating systematic biases due to training on inadequate(ly labeled) datasets, “both supervised and unsupervised learning methods are, on their own [...], poorly suited methods for the raising of ethical machines” (Kaas, 2021; p.56). Hybrid approaches combine top-down and bottom-up techniques (Etzioni & Etzioni, 2017) and have been suggested as the best solution for programming moral AI as they overcome some of the problems of non-hybrid approaches (Song & Yeung, 2022). Such a hybrid approach may, for example, “have rules describing the ethical boundaries but the agent’s goal may need a data-driven approach, or vice-verse” (Rossi & Mattei, 2019; p.9786).

1.2.5 Related research fields and approaches: Subsumption & delimitation

To briefly elucidate how this dissertation contributes to ongoing scholarly debates, this section situates the investigations of this dissertation within established ‘domains of ethics’ (compare Sætra & Danaher, 2022) and introduces related methods such as ‘value-sensitive design’ (Friedman, 1996). Figure 4 illustrates an overview of the various domains of ethics that are explored in this dissertation.

Pertinent domains of ethics. As mentioned earlier, ethics is positioned in the discipline of philosophy and deals with, amongst others, questions about what is morally right or wrong, what goods or values ought to be promoted, what (actions) should be permitted or forbidden, and what relevant theories and arguments guide these determinations (IEEE, 2021; Nyholm, 2023c). Applied ethics is a branch of ethics focusing on concrete moral issues in a specific domain (Copp, 2005). For example, technology ethics and, more concretely AI ethics, are emerging fields within applied ethics (Waelen, 2022). “Technology ethics is the highest-level technology-exclusive form of applied ethics” (Sætra & Danaher, 2022; p.93), which addresses the use and implications of technology on society and contains lower-level applied technology ethics such as AI ethics, robot

ethics, or machine ethics⁸ (Sætra & Danaher, 2022). AI ethics deals the design, use, and governance of autonomous systems (Gordon & Nyholm, 2021) and its impact on society (Coeckelbergh, 2020). In this context, the concept of ‘Responsible AI’ has been introduced, representing the notion of “taking steps to ensure that AI systems have an acceptably low risk of harming their users or society and, ideally, to increase their likelihood of being socially beneficial” (Askill et al., 2019; p.2). Similarly, ‘Ethical AI’ pertains to the creation, deployment, and use of AI that ensures adherence to ethical principles and related core values (Brey & Dainow, 2021). Another related term is ‘Embedded Ethics’, which refers, amongst others, to the programming of ethical principles into algorithms and – in a broad sense – to the integration of ethical considerations throughout a technology’s entire development process (e.g., in the planning, designing, programming, piloting or testing phase) (McLennan et al., 2022). To achieve embedded ethics and responsible or ethical AI, three integrative components have been identified: ‘Ethics in Design’, ‘Ethics for Design’, and ‘Ethics by Design’ (Dignum, 2018). ‘Ethics in Design’ refers to “the regulatory and engineering methods that support the analysis and evaluation of the ethical implications of AI systems”, while ‘Ethics for Design(ers)’ encompass “the codes of conduct, standards and certification processes that ensure the integrity of developers and users as they research, design, construct, employ and manage artificial intelligent systems” (Dignum, 2018; p.2). ‘Ethics by Design’ aims to technically achieve ethical behavior and decision-making of the AI systems themselves by, for example, letting AI system use ethical principles in their reasoning process or setting suitable constraints to system behavior (Dignum, 2018).

In a more specified form of AI ethics, robot ethics deals with the same issues but exclusively for embodied AI systems, such as SDVs. Closely linked to this field is machine ethics, which – as indicated earlier – is about the ethical behavior of the systems themselves and is less directed at human action in this endeavor (Sætra & Danaher, 2022). “Unlike robot ethics, machine ethics focuses more on embedding ethics into autonomous intelligent systems to allow them to make ethical decisions” (Segun, 2021; p.268). Therefore, machine ethics more closely corresponds with computational ethics and ‘Ethics by Design’ approaches rather than, for example, ‘Ethics for Design’ approaches, which concentrate on the humans behind the technology who shape the operational morality of the AI systems. Pertinent inquiries relate to how to (technically) ensure that AI systems behave ethically and possess ethical reasoning capabilities (Dignum, 2018). Another branch of machine ethics is machine metaethics, which examines the field of machine

⁸ Of course, technology ethics also encompasses other lower-level domains such as internet ethics, which involves exploring the societal implications of internet-based technology (e.g., social media) (Sætra & Danaher, 2022). However, these other domains will not be further expanded upon here, as they are not the focus of this dissertation.

ethics itself by, for example, exploring the ultimate goal or implications of machine ethics (Anderson, 2011; Cave et al., 2018).

Within the realms of robot ethics and machine ethics, various specialized branches can focus on specific AI applications, such as those in education, medicine, or transportation (Sætra & Danaher, 2022). Autonomous driving ethics is a branch specific to the ethics of self-driving vehicles, which aims at practically integrating ethical considerations into (the development process of) SDVs (ISO, 2023). As illustrated in Figure 4, autonomous driving ethics can fall under the categories of both robot ethics and machine ethics. This is because investigations within autonomous driving ethics cover a wide range of topics, including the societal and environmental impact of SDVs, issues related to explainability, data governance/privacy and liability, or the explicit programming of SDV decision-making/behavior (Poszler & Geisslinger, 2021). For example, a few scholars in the field have studied how SDVs will improve public mobility and traffic flow (e.g., more inclusive mobility solutions, less congestion) (e.g., Kassens-Noor et al., 2021; Metz, 2018). Other scholars examine how SDVs may contribute to the misuse of private information or where liability should reside in case of an accident (e.g., Collingwood, 2017). Scholarly work in explainable autonomous driving specializes in, amongst others, building “techniques that are transparent enough to support meaningful interpretations to their intended audience” (Omeiza et al., 2021; p.10143). Academics focusing on decision-making research explore how to program the rationale that SDVs use to select the optimal trajectory from available options (see Essay II in Chapter 3 |). Certainly, these various topics are somewhat interconnected. For example, when it comes to the link between data governance/privacy and the decision-making of SDVs: To be able to identify and make the safest decisions, SDVs will need to collect and process a vast amount of (personal) data in the first place (Future of Privacy Forum, 2017). Other scholars have addressed the connection between explainability and decision-making by stressing the importance of transparently elucidating the (logic behind the) driving behavior of SDVs (Henze et al., 2022).

Although all of these topics are important and warrant attention, this dissertation will primarily concentrate on the decision-making of SDVs (as one representative of AMAs). More precisely, this dissertation will investigate how ethical principles can be integrated into AI systems (with a focus on SDVs) and explore subsequent societal implications, thus making specific contributions to the fields of robot ethics, machine ethics, and autonomous driving ethics.

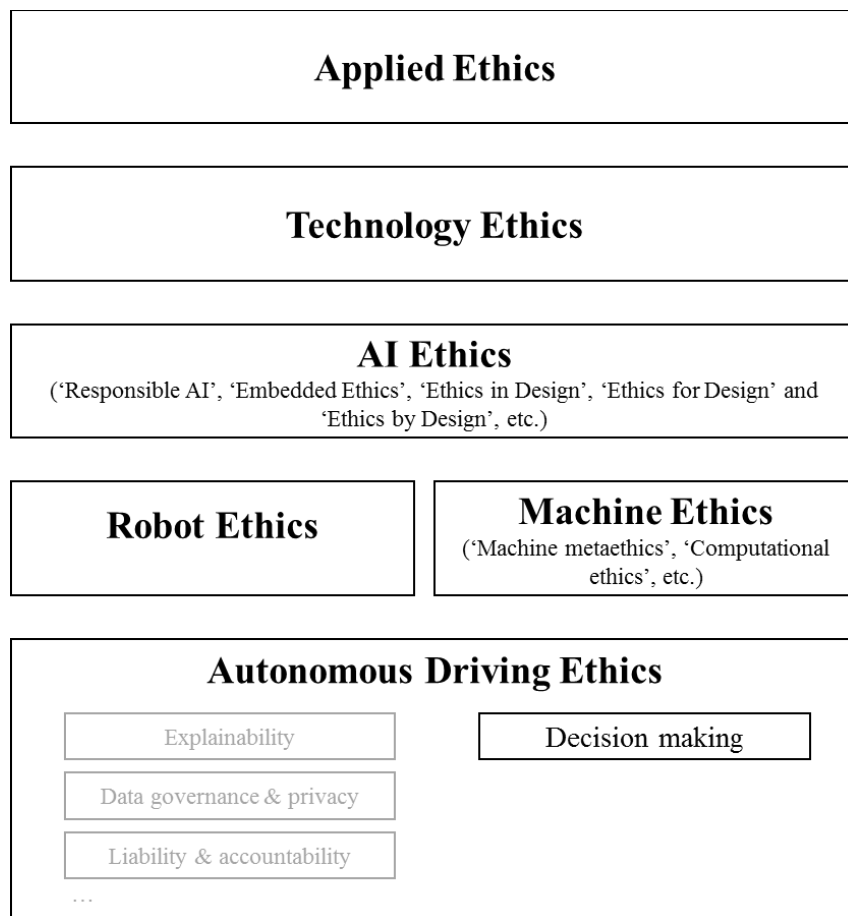


Figure 4: ‘Domains of ethics’ addressed in this dissertation, adapted from Sætra and Danaher (2022)

‘Ethics by Design’ approaches. A prominent method of ‘moralizing technology’ or ‘Ethics by Design’ is value-sensitive design (VSD), which aims to systematically account for human values throughout the design process of technologies (Friedman, 1996). “VSD uses an iterative methodology that integrates conceptual, empirical, and technical investigations” (Verbeek, 2011; p.114), which are all independent of each other but interdependent (Winkler & Spiekermann, 2021). First, conceptual investigations involve a philosophically informed conceptual analysis of affected stakeholders, relevant values at stake, and ethical considerations of particular technologies (Cawthorne, 2022; Manders-Huits, 2011). This rather abstract investigation is complemented with empirical investigations. Through various surveying methods such as observations, focus groups, and interviews, the empirical phase aims to understand and capture human-technology interactions between the system and individuals/society (Cawthorne, 2022; Umbrello, 2019). If there is a mismatch between “what people say they want in a design and what they actually care about in practice”, the results from the conceptual phase need to be reevaluated and adjusted (Umbrello & Yampolskiy, 2022; p.316). The last part (i.e., the technical investigations) is directed more specifically to the design and performance of the particular

technology. Namely, it explores how the technology will support or constrain the values identified in the previous investigations, as well as how these can be translated into certain features and functionalities of the technology (Cawthorne, 2022; Manders-Huits, 2011). Overall, the VSD approach “has accumulated two decades of experience in identifying contextually relevant values and incorporating them into engineering systems” and, more recently, has been adopted in contemporary engineering standards such as IEEE 7000TM (Spiekermann, 2021; p.2). This standard adopts a utilitarian, virtue-ethical, and duty-ethical perspective to unveil relevant stakeholder values. In addition to the elicited values by stakeholders, so-called ‘value-leads’ extend these, drawing on their expert opinion, existing ethical guidelines, standards, pertinent regulations, and so forth. After relevant values are explored and prioritized, fitting technical and organizational system requirements can be developed (Spiekermann, 2023; IEEE, 2021).

While this dissertation generally aligns with ideas and investigation methods outlined in VSD and IEEE 7000TM, it depicts notable differences and extends upon these previous efforts (see more in 5.4.1). For example, VSD not only includes the direct installation of values and parameters into the technology to guide their decision-making processes from a moral perspective, but it is also confined to the activities of designers throughout the entire product lifecycle (Salo-Pöntinen, 2021; Verbeek, 2011). Similarly, this dissertation touches upon company activities that will need to be conducted by the humans behind the AI system, such as creating impact assessments or redress and communication mechanisms (see Essay I in Chapter 2 |). However, the primary focus of this dissertation is on one specific aspect of the product (lifecycle): establishing the underlying reasoning process that guides the ethical decisions of AI systems (see Essay II in Chapter 3 | and Essay III in Chapter 4 |).

1.3 Summary of research gaps and research questions

As mentioned in the previous sections and indicated by scholars such as Gabriel (2020) and Nyholm (2023c), two main challenges need to be tackled to achieve AI systems that are aligned with human core values or goals. First, the technical challenge constitutes investigating the issue of “encoding or embedding values or other principles of ethics into technologies in such a way that they reliably do whatever they ought to do” (Nyholm, 2023c; p.82). To be able to do so, it is central to move away from vague and meaningless AI ethics guidelines (Munn, 2022) and explicitly define and operationalize all values, ethical concepts of interest (e.g., ‘harm’, ‘safety’, ‘fairness’) into ethical (decision-making) principles, which AI systems then abide by (Kaas, 2021). Second, the normative challenge involves determining which concrete (mix of) ethical theories and principles AI systems should adhere to in given circumstances (Dyoub et al., 2020). This challenge remains an ongoing debate, especially in the field of SDVs (Nyholm, 2023c). This dissertation will tackle this issue without, however, providing a normative prescription, but instead presenting possibilities of how to do so. In any case, a necessary step to achieve responsible or ethical AI is to think critically about the system’s societal implications and, correspondingly, incorporate relevant values and ethical principles (Fossa & Cheli, 2023).

Overall, this dissertation will address the following research questions:

➤ *Research question 1 (Practical investigation):* How can ethical principles be integrated into AI systems?

RQ 1.1.: How could computational ethics be administered on a company, industry, and academic level? (addressed in Chapter 2 |)

RQ 1.2: What are functional advantages and disadvantages of applying particular ethical theories to the decision-making of SDVs, and how can these be integrated within SDVs? (addressed in Chapter 3 |)

RQ 1.3.: What constitutes an SDV’s decision-making process that aligns with ethical theories, shared principles, and policymakers’ demands? (addressed in Chapter 4 |)

➤ *Research question 2 (Societal investigation):* What are societal implications of integrating ethical principles into AI systems?

RQ 2.1.: What are reasons for and against computational ethics and resulting artificial moral agents? (addressed in Chapter 2 |)

RQ 2.2.: What are social and moral/legal advantages and disadvantages of applying ethical theories to the decision-making of SDVs? (addressed in Chapter 3 |)

While the subordinated research questions will be answered in individual chapters, answers to the two main research questions (i.e., with respect to the practical investigation and societal investigation) will be summarized in Chapter 5 | (5.2 and 5.3 specifically).

1.4 Research methods

Given the novelty of the topic of inquiry (i.e., integrating ethical principles into AI systems), this dissertation adopts an exploratory research approach (Richey & Klein, 2014). In doing so, the essays in this dissertation follow different methodological, exploratory approaches. Essay I is based on a qualitative expert study, and Essay II constitutes a systematic literature review. With these two essays, this dissertation combines primary data (i.e., insights from the qualitative expert interviews that were independently conducted) and secondary research (i.e., synthesis of prior publications). In addition, derived from the findings of the exploratory phase of this dissertation, Essay III proposes a detailed conceptual framework (i.e., a concrete ethical decision-making process for SDVs). Considering this final output, it can be traced how the topic of inquiry in this dissertation became more focused from an investigation of formalizing ethical principles within AI systems generally (Essay I) towards a more fine-grained investigation of how to do so in the context of SDVs (Essay II and Essay III).

Self-driving vehicles represent a fitting use case for this dissertation's investigation. For one, the relationship between transportation and core values has always been intertwined, with safety as a pivotal value facing potential jeopardy (Fossa & Cheli, 2023). Namely, every year, approximately 1.19 million people die as a result of a traffic accident (WHO, 2023). It is assumed that SDVs could reduce the number of road traffic deaths by 90% (Fagnant & Kockelman, 2015). After all, they “cannot drink and drive, text and drive, or drive aggressively or distractedly” (Evans, 2021; p.316). Thus, scholars have even started to discuss to what extent compelling moral arguments exist against human drivers and whether legislation should prohibit them (Sparrow & Howard, 2017; Müller & Gogoll, 2020). While it is questionable and unlikely that human driving will ever be forbidden (Ratoff, 2022), it is postulated that SDVs will be the first type of robot that will enter society at a mass scale (Lin et al., 2017). In fact, SDVs are already running on the streets today (Edward, 2023), taking over many decisions in traffic. They decide when to brake, how much distance to keep from others, and in most extreme cases, they need to decide in imminent crashes. After all, examples like the Uber crash in 2018 have shown that SDVs can still lead to fatal accidents (McGee, 2019). Thus, SDV decisions are of normative significance (Dietrich & Weisswange, 2019) and need to be sensitive to moral reasons (Nyholm, 2023c).

This turns SDVs into an exemplary case of artificial moral agents. Similarly, given that SDVs operate in transportation and can endanger lives, SDVs – especially the vehicles' components

responsible for scanning the environment or the trajectory planning (Löfling, 2023) – can be classified into ‘high-risk’ AI systems. High-risk AI systems are characterized by their potential to “create a high risk to the health and safety or fundamental rights of natural persons” (European Commission, 2021; p.13), warranting the fulfillment of strict obligations (e.g., adequate risk assessment and mitigation systems) before their introduction to the market. Therefore, a legal viewpoint further underpins the importance of investigating this dissertation’s topic in the context of SDVs. Lastly, road traffic presents a relatively straightforward environment, and SDVs seem to be a practicable technological application for the research subject in this dissertation. This is because road traffic is relatively organized compared to AI systems in other contexts, such as autonomous weapons in warfare. Involved parties must follow the same traffic rules and typically cooperate toward common goals (Nyholm, 2023b). Overall, because of the core values at stake in transportation (i.e., safety), the (predicted) advanced level of SDV adoption, SDVs’ suitable classification into ‘high-risk’ AI systems and AMAs, and the presence of a relatively organized environment (i.e., road traffic), SDVs represent a fitting technological application that will be consulted as a key use case throughout this dissertation.

The research process of this dissertation is illustrated in Figure 5, and the utilized method of each essay is summarized in the following. Overall, this dissertation aims to build theory and lay the groundwork for future research in the respective field in that, for example, the proposed conceptual models can be tested in quantitative-empirical studies (Liu & Stephens, 2019) (see 5.4.2).

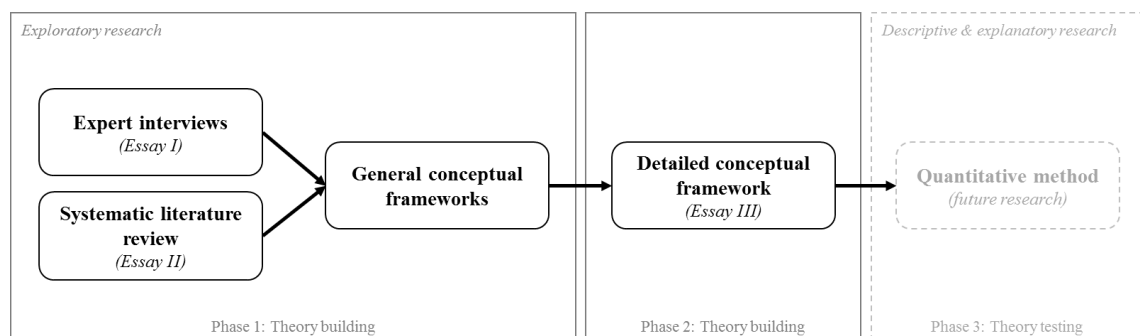


Figure 5: Research process and methods, adapted from Liu and Stephens (2019)

Essay I. The study in Essay I (Chapter 2 |) is based on interviews with twelve experts from disciplines such as philosophy, AI and cognitive sciences, who are knowledgeable about computational ethics and AMAs. Given the infancy of research investigating the topic at hand, qualitative expert interviews are a suitable research method to gather knowledge in this relatively unexplored field (Bogner et al., 2014). Indeed, such research methodologies have already been

used in previous research on AI ethics and computational ethics (e.g., Jacobs & IJsselstein, 2021; Portmann & D'Onofrio, 2022). The interviewed experts were identified via Google Scholar and the snowball sampling technique (Goldstein, 2002).

The interview guide was semi-structured and centered on the expert's personal experiences with computational ethics and resulting technologies (i.e., AMAs), reasons for and against them, as well as their implications and implementation. The interviews were conducted via Zoom between November 2022 and February 2023. The recruitment and interviewing process of additional experts stopped when data saturation was reached (Guest et al., 2020). The interview transcripts were then analyzed using the software MaxQDA by drawing on a manual, inductive coding methodology (Corbin & Strauss, 1994). In addition, AI Assist, the automatic code generation feature of MaxQDA, was utilized to supplement the manually established codes. Ultimately, all identified codes, themes, and dimensions were synthesized into a data structure (Gioia et al., 2012), presenting an overall model for facilitating computational ethics and AMAs.

Data validity was ensured throughout the study. Before data collection, the interview guide was discussed and pretested with experts in qualitative research and representatives from the target sample. Throughout the data analysis process, coding results and interpretations were crosschecked by the co-authors, external researchers (in line with Silverman, 2015), and the automatically generated codes from AI Assist, aiming to enhance objectivity. Preliminary versions of this study were reviewed and discussed with the twelve interviewed experts and the broader science community (e.g., at international conferences).

Essay II. The methodology employed in the second essay (Chapter 3 |) involves conducting a systematic review of the autonomous driving ethics literature. In particular, publications that addressed the application of ethical theories to SDVs' decision-making processes were examined. A review of this sort helps to synthesize prior research and facilitate theory development and an updated research agenda, thereby serving as "a firm foundation for advancing knowledge" (Webster & Watson, 2002; p.xiii). To do so, this article conducted a three-stage iterative process (adapted from Theurer et al., 2018) consisting of the following steps: identification of the relevant literature, structural, and in-depth content analysis of the literature, and integration of the literature.

For the literature search, the databases ScienceDirect, EBSCO Business Source Premier, Scopus, and Web of Science were consulted. Based on predetermined search terms, 884 publications until July 2023 were found, out of which 101 journal articles, book chapters, and contributions to conference proceedings were identified as relevant publications and subjected to further analysis.

The structural analysis addressed a synthesis of the formal publication criteria such as the paper's year of publication, its discipline, or utilized methodology. For the content analysis, an inductive coding and clustering process was manually conducted via the MaxQDA software to allow themes/codes to emerge bottom-up from the identified publications themselves (Fereday & Muir-Cochrane, 2006). In line with Gioia et al. (2012), 1st-order codes were organized into 2nd-order themes, which, in turn, were distilled into three aggregated dimensions (i.e., 'identified ethical theories' that were mentioned in reference to ethical decision-making of SDVs; 'advantages and disadvantages of the identified theories' when applying these to the decision-making of SDVs; 'suggested solutions' for integrated ethical decision-making frameworks and algorithms of SDVs).

Essay III. While the initial two essays establish broader conceptual frameworks, demonstrating how to facilitate the development of AMAs and integrate ethical principles into SDVs, Essay III (Chapter 4 |) expands upon these frameworks by delving into more comprehensive and precise theory building. In terms of comprehensiveness, Essay III incorporates the earlier generated insights from scholarly investigations with further (ethical) considerations drawn from established engineering standards, ethical guidelines, and regulations. Contemplating these documents is crucial, given AI systems such as SDVs are expected to be ethical, lawful, and robust (HLEG, 2019). As this essay – in addition to applying traditional ethical theories – utilizes ethical guidelines previously compiled by teams of experts as resources, it partly aligns with the 'ethics by committee' method (Nyholm, 2023c).

In terms of precision, Essay III advances the upstream theory-building efforts (especially the ones performed in Essay II) by offering a more practice-oriented output. Namely, this essay establishes system requirements and a detailed conceptual, partly mathematical framework in the form of a decision-making process for SDVs facing hazardous situations in traffic. The process is elaborated with a hypothetical, simplified traffic scenario and depicts the underlying calculations for each of the five decision-making steps.

1.5 Dissertation structure and summary of the three essays

To answer the research questions indicated in Section 1.3, this doctoral dissertation is structured as follows. After this introductory chapter, Chapter 2 | through 4 | form a trio of essays aimed at addressing the research questions. A concise overview of the research objectives and key results of the three essays is illustrated in Table 1. The essays in this dissertation are highly interconnected, with each essay building upon the identified open questions and findings of the preceding one(s), thus gradually adding components to the overall scientific inquiry. To elaborate, Chapter 2 | investigates the topic of integrating ethical principles within AI systems without focusing on a specific technological application, but rather, it centers on the method of computational ethics and AMAs in general. One of the derived recommendations from the expert interviews in Chapter 2 | is that the development of AMAs should always be application-specific and accompanied by an ‘ethical principles investigation’. Therefore, Chapter 3 | represents such an ‘ethical principles investigation’ for one specific technological application (i.e., SDVs) by evaluating ethical theories that could be applied to the decision-making logic of SDVs. Based on this preliminary work and additional requirements formulated in contemporary legal/policy drafts, ethical guidelines, and technical standards, Chapter 4 | establishes a precise ethical decision-making model for SDVs during hazardous situations in traffic. Afterward, Chapter 5 | adds a comprehensive review of the research findings derived from the three essays, outlining their implications for society and practitioners. It also highlights how this dissertation contributes to the scientific community, acknowledges its limitations, and points toward directions for future research. Chapter 6 | provides a brief summary of concluding remarks.

Title	Chapter 2	Chapter 3	Chapter 4
	Formalizing Ethical Principles Within AI Systems: Experts' Opinions on Why (Not) And How to do it	Applying Ethical Theories to the Decision-Making of Self-Driving Vehicles: A Systematic Review and Integration of the Literature	Ethical Decision-Making for Self-Driving Vehicles: A Proposed Model & List of Value-Laden Terms That Warrant (Technical) Specification
Research Goals	1) Investigate reasons for and against computational ethics and AMAs 2) Investigate how computational ethics and AMAs could be implemented in practice	1) Investigate advantages and disadvantages of applying traditional ethical theories to SDVs' decision-making 2) Investigate how ethical theories can be integrated into SDVs' ethical decision-making	1) Investigate system requirements for SDVs that align with existing standards, regulatory drafts, and ethical guidelines 2) Investigate how to structure SDVs' decision-making processes in hazardous traffic situations
Theoretical Background	Ethics in design; Ethics by design; Computational ethics	Normative ethics; Autonomous driving ethics	Autonomous driving ethics; Existing ethical guidelines, technical standards, and regulatory efforts concerning SDVs' 'ethical' programming
Research Methodology	Qualitative expert interviews	Systematic literature review	Theoretical, mathematical framework development
Main Findings & Contributions	➤ Indicated arguments for and against computational ethics and AMAs stem from a practical, societal, and epistemic perspective ➤ Implementation recommendations for companies' development, industry's governance, and scientific inquiries are identified ➤ Overarching model is proposed, illustrating key internal activities for companies and external enablers to facilitate the responsible development of AMAs	➤ Each ethical theory can be practically applied to the ethical decision-making of SDVs ➤ The application of each ethical theory to SDVs holds social, moral/legal, and functional advantages and disadvantages ➤ Summarizing model for integrating various ethical theories into SDVs' decision-making is sketched	➤ System requirements for an SDV's ethical decision-making process are summarized ➤ A precise ethical decision-making model for SDVs during hazardous situations is proposed and elaborated in an imaginary traffic scenario ➤ A list of value-laden terms that demand technical specification is provided; exemplary calculations, technical measures/indicators, and theories that could underlie their concretization are presented
Publication Status	Published in: AI and Ethics; https://doi.org/10.1007/s43681-024-00425-6 Presented at e.g.,: 2023 Forum on Philosophy, Engineering, and Technology, Delft	Published in: Technology in Society; https://doi.org/10.1016/j.techsoc.2023.102350 Presented at e.g.,: The 2022 International Society of Business Economics and Ethics World Congress, Bilbao	Published in: Science and Engineering Ethics https://doi.org/10.1007/s11948-024-00513-0 Extended abstract published in: Proceedings of the Conference on Computer Ethics; Presented at e.g.,: International Conference on Computer Ethics: Philosophical Enquiry 2023, Chicago

Table 1: Summary of the essays

References

- Abramson, P., & Inglehar, R. F. (1995). *Value change in global perspective*. The University of Michigan Press.
- Agrawal, A., Gans, J., & Goldfarb, A. (2018). Prediction, judgment, and complexity: a theory of decision-making and artificial intelligence. In A. Agrawal, J. Gans, J., & A. Goldfarb (Eds.), *The economics of artificial intelligence: An agenda* (pp. 89-110). University of Chicago Press.
- Ahmed, H. U., Huang, Y., Lu, P., & Bridgelall, R. (2022). Technology developments and impacts of connected and autonomous vehicles: An overview. *Smart Cities*, 5(1), 382-404. <https://doi.org/10.3390/smartcities5010022>
- Anderson, S. L. (2011). Machine metaethics. In M. Anderson, & S. L. Anderson (Eds.), *Machine ethics* (pp. 21-27). Cambridge University Press.
- Anderson, M., & Anderson, S. L. (2011). *Machine ethics*. Cambridge University Press.
- Askill, A., Brundage, M., & Hadfield, G. (2019). The role of cooperation in responsible AI development. *arXiv preprint arXiv:1907.04534*. <https://doi.org/10.48550/arXiv.1907.04534>
- Awad, E., & Levine, S. (2020). *Why we should crowdsource AI ethics (and how to do so responsibly)*. Retrieved from: <https://behavioralscientist.org/why-we-should-crowdsource-ai-ethics-and-how-to-do-so-responsibly/>
- Awad, E., Levine, S., Anderson, M., Anderson, S. L., Conitzer, V., Crockett, M. J., Everett, J. A., Evgeniou, T., Gopnik, A., Jamison, J. C., Kim, T. W., Liao, S. M., Meyer, M. N., Mikhail, J., Opoku-Agyemang, K., Borg, J. S., Schroeder, J., Sinnott-Armstrong, W., Slavkovik, M., et al. (2022). Computational ethics. *Trends in Cognitive Sciences*, 26(5), 388-405. <https://doi.org/10.1016/j.tics.2022.02.009>
- Bakiner, O. (2023). What do academics say about artificial intelligence ethics? An overview of the scholarship. *AI and Ethics*, 3(2), 513-525. <https://doi.org/10.1007/s43681-022-00182-4>
- Bandura, A. (1991). Social cognitive theory of moral thought and action. In W. M. Kurtines & J. L. Gewirtz (Eds.), *Handbook of moral behavior and development, vol. 1* (pp. 54-103). Erlbaum.
- Beijing Academy of Artificial Intelligence (BAAI) (2019). *Beijing AI Principles*. Retrieved from: <https://www.baai.ac.cn/news/beijing-ai-principles-en.html>
- Beresford, C. (2021). *Honda Legend Sedan with Level 3 Autonomy Available for Lease in Japan*. Retrieved from: <https://www.caranddriver.com/news/a35729591/honda-legend-level-3-autonomy-leases-japan/>
- Bietti, E. (2020). From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 210-219). <https://doi.org/10.1145/3351095.3372860>
- Bin-Nun, A. Y., Derler, P., Mehdipour, N., & Tebbens, R. D. (2022). How should autonomous vehicles drive? Policy, methodological, and social considerations for designing a driver. *Humanities and social sciences communications*, 9(1), 1-13. <https://doi.org/10.1057/s41599-022-01286-2>

BMW Group (2020). *BMW Group code of ethics for artificial intelligence*. Retrieved from: https://www.bmwgroup.com/content/dam/grpw/websites/bmwgroup_com/downloads/ENG_PR_CodeOfEthicsForAI_Short.pdf

Bogner, A., Littig, B., & Menz, W. (2014). *Interviews mit Experten: eine praxisorientierte Einführung*. Springer-Verlag.

Bonnemains, V., Saurel, C., & Tessier, C. (2018). Embedded ethics: some technical and ethical challenges. *Ethics and Information Technology*, 20(1), 41-58. <https://doi.org/10.1007/s10676-018-9444-x>

Brey, P., & Dainow, B. (2021). Ethics by design and ethics of use in AI and robotics. The SIENNA project-Stakeholder-informed ethics for new technologies with high socioeconomic and human rights impact. *SIENNA*. Retrieved from: https://sienna-project.eu/digitalAssets/915/c_915554-1_1-k_sienna-ethics-by-design-and-ethics-of-use.pdf

Bringsjord, S., Arkoudas, K., & Bello, P. (2006). Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems*, 21(4), 38-44. <https://doi.org/10.1109/MIS.2006.82>

Caballero, W. N., Rios Insua, D., & Banks, D. (2023). Decision support issues in automated driving systems. *International Transactions in Operational Research*, 30(3), 1216-1244. <https://doi.org/10.1111/itor.12936>

Campbell, R. (2017). Learning from moral inconsistency. *Cognition*, 167, 46-57. <https://doi.org/10.1016/j.cognition.2017.05.006>

Carlson, D. S., Kacmar, K. M., & Wadsworth, L. L. (2009). The impact of moral intensity dimensions on ethical decision-making: Assessing the relevance of orientation. *Journal of Managerial Issues*, 21(4), 534-551.

Cave, S., Nyrop, R., Vold, K., & Weller, A. (2018). Motivations and risks of machine ethics. *Proceedings of the IEEE*, 107(3), 562-574. <https://doi.org/10.1109/JPROC.2018.2865996>

Cawthorne, D. (2022). Robot Ethics: Ethical Design Considerations. In D. Herath, & D. St-Onge (Eds.), *Foundations of Robotics: A Multidisciplinary Approach with Python and ROS* (pp. 473-491). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-1983-1_16

Coeckelbergh, M. (2020). *AI ethics*. The MIT Press.

Collingwood, L. (2017). Privacy implications and liability issues of autonomous vehicles. *Information & Communications Technology Law*, 26(1), 32-45. <https://doi.org/10.1080/13600834.2017.1269871>

Copp, D. (2005). *The Oxford handbook of ethical theory*. Oxford University Press.

Corbin, J., & Strauss, A. (1994). Grounded theory methodology. *Handbook of qualitative research*, 17, 273-285.

Coskun, S. (2021). Autonomous overtaking in highways: A receding horizon trajectory generator with embedded safety feature. *Engineering Science and Technology, an International Journal*, 24(5), 1049-1058. <https://doi.org/10.1016/j.jestch.2021.02.005>

Danks, D. (2022). Digital ethics as translational ethics. In I. Vasiliu-Feltes, & J. Thomason (Eds.), *Applied ethics in a digital world* (pp. 1-15). IGI Global. <https://doi.org/10.4018/978-1-7998-8467-5.ch001>

Deemantha, R. G. S., & Hettige, B. (2019). Autonomous Car: Current Issues, Challenges and Solution: A Review. In *Proceedings of the 15th International Research Conference, Beijing, China* (pp. 1-6).

Dietrich, M., & Weisswange, T. H. (2019). Distributive justice as an ethical principle for autonomous vehicle behavior beyond hazard scenarios. *Ethics and Information Technology*, 21(3), 227-239. <https://doi.org/10.1007/s10676-019-09504-3>

Dignum, V. (2018). Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology*, 20(1), 1-3. <https://doi.org/10.1007/s10676-018-9450-z>

Dolgov, D. and Urmson, C. (2014). *Controlling vehicle lateral lane positioning*. Retrieved from: <https://patents.google.com/patent/US8781670B2/en>

Dyoub, A., Costantini, S., Lisi, F. A., & Letteri, I. (2020). Logic-based Machine Learning for Transparent Ethical Agents. In *CILC* (pp. 169-183).

Edward, K. (2023). Mercedes-Benz First To Gain U.S. Approval For Level 3 Automated Driving System. *Forbes*. Retrieved from: <https://www.forbes.com/sites/kyleedward/2023/09/28/mercedes-benz-first-to-gain-us-approval-for-level-3-automated-driving-system/?sh=59e249065c8e>

Etienne, H. (2022). A practical role-based approach for autonomous vehicle moral dilemmas. *Big Data & Society*, 9(2), 20539517221123305. <https://doi.org/10.1177/2053951722112333>

ETSI (2019). *Intelligent Transport System (ITS); Vulnerable Road Users (VRU) awareness; Part 1: Use Cases definition; Release 2*. Retrieved from: https://www.etsi.org/deliver/etsi_tr/103300_103399/10330001/02.01.01_60/tr_10330001v020101p.pdf

Etzioni, A., & Etzioni, O. (2017). Incorporating Ethics into Artificial Intelligence. *The Journal of Ethics*, 21(4), 403-418. <https://doi.org/10.1007/s10892-017-9252-2>

European Commission, Directorate-General for Research and Innovation (2020). *Ethics of connected and automated vehicles: Recommendations on road safety, privacy, fairness, explainability and responsibility*. Retrieved from: <https://op.europa.eu/en/publication-detail/-/publication/89624e2c-f98c-11ea-b44f-01aa75ed71a1/language-en/format-PDF/source-search>

European Commission (2021). *Proposal for a Regulation of the European Parliament and of the Council. Laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative acts*. Retrieved from: <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206>

European Commission (2022). *Commission implementing regulation (EU) 2022/1426*. Retrieved from: <https://eurlex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R1426>

Evans, K. (2021). *The Implementation of Ethical Decision Procedures in Autonomous Systems: The Case of the Autonomous Vehicle*. Doctoral dissertation, Sorbonne université.

Evans, K., de Moura, N., Chauvier, S., & Chatila, R. (2023). Automated Driving Without Ethics: Meaning, Design and Real-World Implementation. In F. Fossa, & F. Cheli (Eds.), *Connected and Automated Vehicles: Integrating Engineering and Ethics* (pp. 123-143). Springer Nature Switzerland.

Fagnant, D. J., & Kockelman, K. (2015). Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy and Practice*, 77, 167-181. <https://doi.org/10.1016/j.tra.2015.04.003>

Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International journal of qualitative methods*, 5(1), 80-92. <https://doi.org/10.1177/16094069060050010>

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center for Internet & Society*. Retrieved from: <https://dash.harvard.edu/handle/1/42160420>

Flipse, S. M., & Puylaert, S. (2018). Organizing a collaborative development of technological design requirements using a constructive dialogue on value profiles: A case in automated vehicle development. *Science and engineering ethics*, 24, 49-72. <https://doi.org/10.1007/s11948-017-9877-3>

Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Lütge, C., Madelin, R., Pagallo, U., Rossi, F., et al. (2018). An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28, 689-707. <https://doi.org/10.1007/s11023-018-9482-5>

Fossa, F. (2023). Unavoidable Collisions. The Automation of Moral Judgment. In F. Fossa (Ed.), *Ethics of Driving Automation: Artificial Agency and Human Values* (pp. 65-94). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-22982-4_4

Fossa, F., & Cheli, F. (2023). *Connected and Automated Vehicles: Integrating Engineering and Ethics* (Vol. 67). Springer Nature. <https://doi.org/10.1007/978-3-031-39991-6>

Friedman, B. (1996). Value-sensitive design. *interactions*, 3(6), 16-23.

Friedman, B., Kahn, P. H., Borning, A., & Hultgren, A. (2013). Value sensitive design and information systems. In N. Doorn, D. Schuurbiers, I. van de Poel, & M. E. Gorman (Eds.), *Early engagement and new technologies: Opening up the laboratory* (pp. 55-95). Springer Science & Business Media.

Future of Privacy Forum (2017). *Data and the connected car*. Retrieved from https://fpf.org/wp-content/uploads/2017/06/2017_0627-FPF-Connected-Car-Infographic-Version-1.0.pdf

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and machines*, 30(3), 411-437. <https://doi.org/10.1007/s11023-020-09539-2>

Geisslinger, M., Poszler, F., Betz, J., Lütge, C., & Lienkamp, M. (2021). Autonomous driving ethics: From trolley problem to ethics of risk. *Philosophy & Technology*, 34(4), 1033-1055. <https://doi.org/10.1007/s13347-021-00449-4>

Gentzel, M. (2020). Classical liberalism, discrimination, and the problem of autonomous cars. *Science and Engineering Ethics*, 26(2), 931-946. <https://doi.org/10.1007/s11948-019-00155-7>

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2012). Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology. *Organizational research methods*, 16 (1), 15–31. <https://doi.org/10.1177/1094428112452151>

Gogoll, J., Zuber, N., Kacianka, S., Greger, T., Pretschner, A., & Nida-Rümelin, J. (2021). Ethics in the software development process: from codes of conduct to ethical deliberation. *Philosophy & Technology*, 34(4), 1085-1108. <https://doi.org/10.1007/s13347-021-00451-w>

Goldstein, K. (2002). Getting in the door: Sampling and completing elite interviews. *PS: Political Science and Politics*, 35(4), 669-672. <https://doi.org/10.1017/S1049096502001130>

Goodall, N. J. (2014). Ethical decision making during automated vehicle crashes. *Transportation Research Record: Journal of the Transportation Research Board*, 2424(1), 58–65. <https://doi.org/10.3141/2424-07>

Gordon, J.-S., & Nyholm, S. (2021). Ethics of artificial intelligence. *Internet Encyclopedia of Philosophy*. Retrieved from: <https://iep.utm.edu/ethics-of-artificial-intelligence/>

Gryz, J. (2020). Some Technical Challenges in Designing an Artificial Moral Agent. In L. Rutkowski, R. Scherer, M. Korytkowski, W. Pedrycz, R. Tadeusiewicz, & J. M. Zurada (Eds.), *Artificial Intelligence and Soft Computing. ICAISC 2020. Lecture Notes in Computer Science, vol 12416* (pp. 481-491). Springer International Publishing. https://doi.org/10.1007/978-3-030-61534-5_43

Guest G., Namey, E., & Chen, M. (2020). A simple method to assess and report thematic saturation in qualitative research. *PLoS One*, 15(5), e0232076. <https://doi.org/10.1371/journal.pone.0232076>

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California management review*, 61(4), 5-14. <https://doi.org/10.1177/00081256198649>

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4), 814–834. <https://doi.org/10.1037/0033-295X.108.4.814>

Henze, F., Faßbender, D., & Stiller, C. (2022). How Can Automated Vehicles Explain Their Driving Decisions? Generating Clarifying Summaries Automatically. In *2022 IEEE Intelligent Vehicles Symposium (IV)* (pp. 935-942). IEEE. <https://doi.org/10.1109/IV51971.2022.9827197>

High-level Expert Group on Artificial Intelligence (HLEG) (2019). *Ethics Guidelines for trustworthy AI*. Retrieved from: <https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence>

Hoffmann, C. H. N. (2021). On formal ethics versus inclusive moral deliberation. *AI and Ethics*, 1, 313-329. <https://doi.org/10.1007/s43681-021-00045-4>

Hull, D. (2015). Tesla Starts Rolling Out Autopilot Features. *Bloomberg*. Retrieved from: <https://www.bloomberg.com/news/articles/2015-10-14/tesla-software-upgrade-adds-automated-lane-changing-to-model-s?embedded-checkout=true>

Hunyadi, M. (2019). Artificial moral agents. Really?. In J.-P. Laumond, E. Danblon, & C. Pieters (Eds.), *Wording Robotics: Discourses and Representations on Robotics* (pp. 59-69). Springer. https://doi.org/10.1007/978-3-030-17974-8_5

IEEE (2019). *Ethically aligned design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. Available: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

IEEE (2021). *IEEE Std 7000TM-2021: IEEE Standard Model Process for Addressing Ethical Concerns during System Design*. Retrieved from: <https://engagestandards.ieee.org/ieee-7000-2021-for-systems-design-ethical-concerns.html>

ISO (2022). *ISO/IEC 22989:2022 - Information technology, Artificial intelligence. Artificial intelligence concepts and terminology*. Retrieved from: <https://www.iso.org/standard/74296.html>

ISO (2023). *ISO 39003:2023 - Road traffic safety (RTS): Guidance on ethical considerations relating to safety for autonomous vehicles*. Retrieved from: <https://www.iso.org/obp/ui/fr/#iso:std:iso:39003:dis:ed-1:v1:en:fig:4>

Jacobs, N., & IJsselsteijn, W. (2021). Bridging the Theory-Practice Gap: Design-Experts on Capability Sensitive Design. *International Journal of Technoethics*, 12(2), 1-16. <https://doi.org/10.4018/IJT.2021070101>

Jones, T. M. (1991). Ethical decision making by individuals in organizations: An issue-contingent model. *Academy of management review*, 16(2), 366-395. <https://doi.org/10.5465/amr.1991.4278958>

Kaas, M. H. (2021). Raising Ethical Machines: Bottom-Up Methods to Implementing Machine Ethics. In S. J. Thompson (Ed.), *Machine Law, Ethics, and Morality in the Age of Artificial Intelligence* (pp. 47-68). IGI Global. <https://doi.org/10.4018/978-1-7998-4894-3.ch004>

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *American psychologist*, 58(9), 697-720. <https://doi.org/10.1037/0003-066X.58.9.697>

Kassens-Noor, E., Cai, M., Kotval-Karamchandani, Z., & Decaminada, T. (2021). Autonomous vehicles and mobility for people with special needs. *Transportation research part A: policy and practice*, 150, 385-397. <https://doi.org/10.1016/j.tra.2021.06.014>

Kim, T. W., Hooker, J., & Donaldson, T. (2020). Taking Principles Seriously: A Hybrid Approach to Value Alignment. *arXiv preprint arXiv:2012.11705*. <https://doi.org/10.48550/arXiv.2012.11705>

Kim, T. W., Hooker, J., & Donaldson, T. (2021). Taking principles seriously: A hybrid approach to value alignment in artificial intelligence. *Journal of Artificial Intelligence Research*, 70, 871-890. <https://doi.org/10.1613/jair.1.12481>

Koopman, P., & Widen, W. H. (2023). Ethical Design and Testing of Automated Driving Features. *University of Miami School of Law*, 4336441. <http://dx.doi.org/10.2139/ssrn.4336441>

Kosuru, V. S. R., & Venkitaraman, A. K. (2023). Advancements and challenges in achieving fully autonomous self-driving vehicles. *World Journal of Advanced Research and Reviews*, 18(1), 161-167. <https://doi.org/10.30574/wjarr.2023.18.1.0568>

LaCroix, T., & Luccioni, A. S. (2022). Metaethical perspectives on 'Benchmarking' AI ethics. *arXiv preprint arXiv:2204.05151*. <https://doi.org/10.48550/arXiv.2204.05151>

Lin, P., Abney, K., & Jenkins, R. (2017). *Robot ethics 2.0: From autonomous cars to artificial intelligence*. Oxford University Press.

Liu, Z., & Stephens, V. (2019). Exploring innovation ecosystem from the perspective of sustainability: Towards a conceptual framework. *Journal of Open Innovation: Technology, Market, and Complexity*, 5(3), 48. <https://doi.org/10.3390/joitmc5030048>

Löfling, N. (2023). Impact of the EU's AI Act proposal on automated and autonomous vehicles. *Bird & Bird*. Retrieved from: <https://www.twobirds.com/en/insights/2023/global/impact-of-the-eus-ai-act-proposal-on-automated-and-autonomous-vehicles>

Lucifora, C., Grasso, G. M., Perconti, P., & Plebe, A. (2021). Moral reasoning and automatic risk reaction during driving. *Cognition, Technology & Work*, 23(4), 705-713. <https://doi.org/10.1007/s10111-021-00675-y>

Lütge, C. (2017). The German ethics code for automated and connected driving. *Philosophy & Technology*, 30(4), 547-558. <https://doi.org/10.1007/s13347-017-0284-0>

Lütge, C. (2024). *Wirtschaftsethik in realistischer Perspektive*. Mohr Siebeck. <https://doi.org/10.1628/978-3-16-162808-5>

Lütge, C., Poszler, F., Acosta, A. J., Danks, D., Gottehrer, G., Mihet-Popa, L., & Naseer, A. (2021). AI4People: Ethical Guidelines for the Automotive Sector—Fundamental Requirements and Practical Recommendations. *International Journal of Technoethics*, 12(1), 101-125. <https://doi.org/10.4018/IJT.20210101.oa2>

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* (pp. 117-124). <http://dx.doi.org/10.1145/2696454.2696458>

Manders-Huits, N. (2011). What values in design? The challenge of incorporating moral values into design. *Science and engineering ethics*, 17(2), 271-287. <https://doi.org/10.1007/s11948-010-9198-2>

Martinho, A., Kroesen, M., & Chorus, C. (2021). Computer Says I Don't Know: An Empirical Approach to Capture Moral Uncertainty in Artificial Intelligence. *Minds and Machines*, 31(2), 215-237. <https://doi.org/10.1007/s11023-021-09556-9>

McGee, P. (2019). Uber back-up driver faulted in fatal autonomous car crash. *Financial Times*. Retrieved from: <https://www.ft.com/content/6d0c5544-0afb-11ea-bb52-34c8d9dc6d84>

McKinsey & Company (2023). *Autonomous driving's future: Convenient and connected*. Retrieved from: <https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/autonomous-drivings-future-convenient-and-connected>

McLennan, S., Fiske, A., Tigard, D., Müller, R., Haddadin, S., & Buyx, A. (2022). Embedded ethics: a proposal for integrating ethics into the development of medical AI. *BMC Medical Ethics*, 23(1), 6. <https://doi.org/10.1186/s12910-022-00746-3>

McNamara, A., Smith, J., & Murphy-Hill, E. (2018). Does ACM's code of ethics change ethical decision making in software development?. In *Proceedings of the 2018 26th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering* (pp. 729-733). <https://doi.org/10.1145/3236024.3264833>

Mercedes-Benz Group (2024). *Wie Mercedes-Benz Künstliche Intelligenz (KI) einsetzt: Zwei Buchstaben, vier Prinzipien*. Retrieved from: <https://group.mercedes-benz.com/verantwortung/compliance/digital/ki-guidelines.html>

Metz, D. (2018). Developing policy for urban autonomous vehicles: Impact on congestion. *Urban Science*, 2(2), 33. <https://doi.org/10.3390/urbansci2020033>

Mezgár, I., & Váncza, J. (2022). From ethics to standards—A path via responsible AI to cyber-physical production systems. *Annual Reviews in Control*, 53, 391-404. <https://doi.org/10.1016/j.arcontrol.2022.04.002>

Microsoft (2022). *Microsoft Responsible AI Standard (v2) – GENERAL REQUIREMENTS*. Retrieved from: <https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE5cmFl?culture=en-us&country=us>

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature machine intelligence*, 1(11), 501-507. <https://doi.org/10.1038/s42256-019-0114-4>

Montréal Declaration Responsible AI (2018). *Montréal Declaration for a Responsible Development of Artificial Intelligence*. Retrieved from: https://declarationmontreal-iaresponsable.com/wp-content/uploads/2023/04/UdeM_Decl-IA-Resp_LA-Declaration-ENG_WEB_09-07-19.pdf

Moor, J. H. (1995). Is ethics computable?. *Metaphilosophy*, 26(1/2), 1-21.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems*, 21(4), 18-21. <https://doi.org/10.1109/MIS.2006.80>

Mordue, G., Yeung, A., & Wu, F. (2020). The looming challenges of regulating high level autonomous vehicles. *Transportation research part A: policy and practice*, 132, 174-187. <https://doi.org/10.1016/j.tra.2019.11.007>

Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., & Floridi, L. (2021). Operationalising AI ethics: barriers, enablers and next steps. *AI & SOCIETY*, 38, 1-13. <https://doi.org/10.1007/s00146-021-01308-8>

Müller, J. F., & Gogoll, J. (2020). Should manual driving be (eventually) outlawed?. *Science and engineering ethics*, 26(3), 1549-1567. <https://doi.org/10.1007/s11948-020-00190-9>

Munn, L. (2022). The uselessness of AI ethics. *AI and Ethics*, 3, 869-877. <https://doi.org/10.1007/s43681-022-00209-w>

National Conference of Commissioners on Uniform State Laws (2019). *Uniform automated operation of vehicles act - National conference of commissioners on uniform state laws*. Retrieved from: https://oohwstcavworkgroup.blob.core.windows.net/media/Default/documents/infrastructure-systems/Meeting_7/WSTC_AVWG_Infrastructure_Subcommittee_Meeting_7_UniformLawCommissionAVModelBill.pdf

Nath, R., & Sahu, V. (2020). The problem of machine ethics in artificial intelligence. *AI & society*, 35(1), 103-111. <https://doi.org/10.1007/s00146-017-0768-6>

NHTSA (2017). *Automated driving systems 2.0: A vision for safety*. Retrieved from: https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/13069a-ads2.0_090617_v9a_tag.pdf.

Nyholm, S. (2023a). Artificial Intelligence, Ethics of. In M. Sellers, & S. Kirste (Eds.), *Encyclopedia of the Philosophy of Law and Social Philosophy*. Springer. https://doi.org/10.1007/978-94-007-6730-0_1093-1

Nyholm, S. (2023b). Minding the Gap (s): Different Kinds of Responsibility Gaps Related to Autonomous Vehicles and How to Fill Them. In In F. Fossa, & F. Cheli (Eds.), *Connected and Automated Vehicles: Integrating Engineering and Ethics* (pp. 1-18). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-39991-6_1

Nyholm, S. (2023c). *This is technology ethics: An introduction*. John Wiley & Sons.

OECD (2024). *Recommendation of the Council on OECD Legal Instruments Artificial Intelligence* - *OECD/LEGAL/0449*. Retrieved from: <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>

Omeiza, D., Webb, H., Jirotko, M., & Kunze, L. (2021). Explanations in autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(8), 10142-10162. <https://doi.org/10.1109/TITS.2021.3122865>

Pendleton, S. D., Andersen, H., Du, X., Shen, X., Meghjani, M., Eng, Y. H., Rus, D., & Ang, M. H. (2017). Perception, planning, control, and coordination for autonomous vehicles. *Machines*, 5(1), 6. <https://doi.org/10.3390/machines5010006>

Pirtle, Z., Tomblin, D., & Madhavan, G. (2021). Reimagining conceptions of technological and societal progress. In Z. Pirtle, D. Tomblin, & G. Madhavan (Eds.), *Engineering and Philosophy: Reimagining Technology and Social Progress* (pp. 1-21). Springer International Publishing. https://doi.org/10.1007/978-3-030-70099-7_1

Portmann, E., & D'Onofrio, S. (2022). Computational ethics. *HMD Praxis der Wirtschaftsinformatik*, 59(2), 447-467. <https://doi.org/10.1365/s40702-022-00855-y>

Poszler, F., & Geisslinger, M. (2021). *AI and Autonomous Driving: Key ethical considerations*. *IEAI Research Brief*. Retrieved from: https://ieai.sot.tum.de/wp-content/uploads/2021/02/ResearchBrief_February2021_AutonomousVehicles_FINAL.pdf

Ratoff, W. (2022). Self-driving Cars and the Right to Drive. *Philosophy & Technology*, 35(3), 57. <https://doi.org/10.1007/s13347-022-00551-1>

Rest, J. R. (1986). *Moral development: Advances in research and theory*. Praeger.

Richey, R. C., & Klein, J. D. (2014). *Design and development research: Methods, strategies, and issues*. Routledge. <https://doi.org/10.4324/9780203826034>

Roff, H. (2018). The folly of trolleys: Ethical challenges and autonomous vehicles. *The Brookings Institution*. Retrieved from: <https://www.brookings.edu/research/the-folly-of-trolleys-ethical-challenges-and-autonomous-vehicles/>

Rossi, F., & Mattei, N. (2019). Building ethically bounded AI. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33) (pp. 9785-9789). <https://doi.org/10.1609/aaai.v33i01.33019785>

SAE International (2021). *Surface vehicle recommended practice: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles (J3016)*. Retrieved from: https://www.sae.org/standards/content/j3016_202104/

Sætra, H. S., & Danaher, J. (2022). To each technology its own ethics: The problem of ethical proliferation. *Philosophy & Technology*, 35(4), 93. <https://doi.org/10.1007/s13347-022-00591-7>

Salo-Pöntinen, H. (2021). AI Ethics-Critical Reflections on Embedding Ethical Frameworks in AI Technology. In *International Conference on Human-Computer Interaction* (pp.311-329). Springer. https://doi.org/10.1007/978-3-030-77431-8_20

Scheutz, M., & Malle, B. F. (2017). Moral robots. In L. S. M. Johnson, & K. S. Rommelfanger (Eds.), *The Routledge handbook of neuroethics* (pp. 363-377). Routledge.

Schwarz, E. (2023). Cybernetics at war. In A. Gruszczak, & S. Kaempf (Eds.), *Routledge handbook of the future of warfare* (Vol. 490) (pp. 297-307). Routledge.

Sclove, R. (1995). *Democracy and technology*. Guilford Press.

Segun, S. T. (2021). From machine ethics to computational ethics. *AI & SOCIETY*, 36(1), 263-276. <https://doi.org/10.1007/s00146-020-01010-1>

Shneiderman, B. (2020). Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(4), 1-31. <https://doi.org/10.1145/3419764>

Siegel, J., & Pappas, G. (2021). Morals, ethics, and the technology capabilities and limitations of automated and self-driving vehicles. *AI & SOCIETY*, 38(1), 213-226. <https://doi.org/10.1007/s00146-021-01277-y>

Silverman, D. (2015). *Interpreting qualitative data*. Sage.

Singer, P. (2011). *Practical ethics*. Cambridge university press.

Song, F., & Yeung, S. H. F. (2022). A pluralist hybrid model for moral AIs. *AI & SOCIETY*, 1-10. <https://doi.org/10.1007/s00146-022-01601-0>

Sparrow, R., & Howard, M. (2017). When human beings are like drunk robots: driverless vehicles, ethics, and the future of transport. *Transportation Research Part C*. <https://doi.org/10.1016/j.trc.2017.04.014>

Spiekermann, S. (2021). From value-lists to value-based engineering with IEEE 7000™. In *2021 IEEE International Symposium on Technology and Society (ISTAS)* (pp. 1-6). IEEE. <https://doi.org/10.1109/ISTAS52410.2021.9629134>

Spiekermann, S. (2023). *Value-Based Engineering: A Guide to Building Ethical Technology for Humanity*. De Gruyter. <https://doi.org/10.1515/9783110793383>

Srivastava, A. (2019). Sense-Plan-Act in Robotic Applications. <https://doi.org/10.13140/RG.2.2.21308.36481>

Taylor, M. (2016). Self-Driving Mercedes-Benzen Will Prioritize Occupant Safety over Pedestrians. *Car and Driver*. Retrieved from: <https://www.caranddriver.com/news/a15344706/selfdriving-mercedes-will-prioritize-occupant-safety-over-pedestrians/>

Theurer, C. P., Tumasjan, A., Welp, I. M., & Lievens, F. (2018). Employer branding: a brand equity-based literature review and research agenda. *International Journal of Management Reviews*, 20(1), 155-179. <https://doi.org/10.1111/ijmr.12121>

The White House Washington (2022). *Blueprint for an AI Bill of Rights – Making automated systems work for the American people*. Retrieved from: <https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf>

Trovati, M., Johnny, O., Xu, X., & Polatidis, N. (2022). A new model for artificial intuition. In E. Pimenidis, P. Angelov, C. Jayne, A. Papaleonidas, & M. Aydin (Eds.), *Artificial Neural Networks and Machine Learning – ICANN 2022. Lecture Notes in Computer Science*, vol 13529 (pp. 454-465). Springer International Publishing. https://doi.org/10.1007/978-3-031-15919-0_38

TÜV Süd (2023). *Ethical Considerations and Autonomous Vehicles: Why and how ethics matter for the implementation of autonomous driving in Germany and beyond*. Retrieved from: <https://www.tuvsud.com/en/resource-centre/white-papers/ethical-considerations-and-autonomous-vehicles>

Umbrello, S. (2019). Beneficial artificial intelligence coordination by means of a value sensitive design approach. *Big Data and Cognitive Computing*, 3(1), 5. <https://doi.org/10.3390/bdcc3010005>

Umbrello, S., & van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. *AI and Ethics*, 1(3), 283-296. <https://doi.org/10.1007/s43681-021-00038-3>

Umbrello, S., & Yampolskiy, R. V. (2022). Designing AI for explainability and verifiability: a value sensitive design approach to avoid artificial stupidity in autonomous vehicles. *International Journal of Social Robotics*, 14(2), 313-322. <https://doi.org/10.1007/s12369-021-00790-w>

Van de Poel, I. (2009). Values in engineering design. In A. Meijers (Ed.), *Philosophy of technology and engineering sciences* (pp. 973-1006). North-Holland. <https://doi.org/10.1016/B978-0-444-51667-1.50040-9>

Van de Poel, I. (2013). Translating values into design requirements. In D. Michelfelder, N. McCarthy, & D. Goldberg (Eds.), *Philosophy and Engineering: Reflections on Practice, Principles and Process. Philosophy of Engineering and Technology*, vol 15 (pp. 253-266). https://doi.org/10.1007/978-94-007-7762-0_20

Van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. *Minds and Machines*, 30(3), 385-409. <https://doi.org/10.1007/s11023-020-09537-4>

Verbeek, P. P. (2011). *Moralizing technology: Understanding and designing the morality of things*. University of Chicago press.

Waelen, R. (2022). Why AI ethics is a critical theory. *Philosophy & Technology*, 35(1), 9. <https://doi.org/10.1007/s13347-022-00507-5>

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Wang, Y., Liu, Z., Zuo, Z., Li, Z., Wang, L., & Luo, X. (2019). Trajectory planning and safety assessment of autonomous vehicles based on motion prediction and model predictive control. *IEEE Transactions on Vehicular Technology*, 68(9), 8546-8556. <https://doi.org/10.1109/TVT.2019.2930684>

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS quarterly*, 26(2), xiii-xxiii.

Wei, J., Snider, J. M., Gu, T., Dolan, J. M., & Litkouhi, B. (2014). A behavioral planning framework for autonomous driving. In *2014 IEEE Intelligent Vehicles Symposium Proceedings* (pp. 458-464). IEEE. <https://doi.org/10.3390/electronics7060084>

Winkler, T., & Spiekermann, S. (2021). Twenty years of value sensitive design: a review of methodological practices in VSD projects. *Ethics and Information Technology*, 23, 17-21. <https://doi.org/10.1007/s10676-018-9476-2>

Woodgate, J., & Ajmeri, N. (2022). Principles for Macro Ethics of Sociotechnical Systems: Taxonomy and Future Directions. *arXiv preprint arXiv:2208.12616*. <https://doi.org/10.48550/arXiv.2208.12616>

World Health Organization (WHO) (2023). *Road traffic injuries*. Retrieved from: <https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries>

Zollo, L., Pellegrini, M. M., & Ciappei, C. (2017). What sparks ethical decision making? The interplay between moral intuition and moral reasoning: lessons from the scholastic doctrine. *Journal of Business Ethics*, 145, 681-700. <https://doi.org/10.1007/s10551-016-3221-8>

Zoshak, J., & Dew, K. (2021). Beyond kant and bentham: How ethical theories are being used in artificial moral agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-15). <https://doi.org/10.1145/3411764.3445102>

2 | Essay I: Formalizing Ethical Principles Within AI Systems: Experts' Opinions on Why (Not) And How to do it

Abstract

AI systems are increasingly put into contexts where computed decisions must be guided by ethical considerations. To develop ethically grounded algorithms and technologies, scholars have suggested computational ethics as an essential frontier, which aims to translate ethical principles into computer code. However, computational ethics has received little attention in academic literature so far, with existing work mainly focusing on its technical implementation, while many open questions concerning its (societal and ethical) implications still need to be resolved. Therefore, in this study, we interviewed twelve experts from philosophy, AI, and cognitive sciences to shed light on computational ethics beyond a technical perspective. Findings suggest that indicated supporting and opposing arguments can be clustered into pragmatic/practical, societal, and epistemic reasons, all of which need to be contemplated when engaging in computational ethics and developing resulting artificial moral agents. Furthermore, the mentioned recommendations for companies' technological design and development, for industry's governance measures, and academia's research endeavors are recapitulated and summarized in a holistic framework that aims to facilitate a reflected implementation of 'ethics in and by design' in the future.

Keywords: *artificial moral agents, computational ethics, ethics by design, ethical decision-making*

Note: This chapter is based on a published paper co-authored by Edy Portmann and Christoph Lütge. Therefore, the plural instead of the singular is used throughout this chapter. Author contributions to this paper and copyright information are summarized in Appendix A: Reference & copyright information by the publisher for the first essay (Essay I, Chapter 2) and Appendix D: Author contributions to the three essays in this dissertation.

Current publication status:

Poszler, F., Portmann, E., & Lütge, C. (2024). Formalizing Ethical Principles Within AI Systems: Experts' Opinions on Why (Not) And How to do it. *AI and Ethics*, 1-29.
<https://doi.org/10.1007/s43681-024-00425-6>

Conference presentations (of previous versions):

Poszler, F., Portmann, E., & Lütge, C. (April, 2023). Translating ethical principles into computer codes: Experts' opinions on why (not) and how to do it. Presented at the 2023 Forum on Philosophy, Engineering, and Technology, Delft.

2.1 Introduction

Advances in artificial intelligence (AI) have increased the range of decisions autonomous machines can take over (Phillips-Wren, 2012). Some of these “machines equipped with automated decisions functions, are [even] intended to be put in contexts where computed decisions have to be guided by ethical considerations“ (Bonnemains et al., 2018; pp.3-4) and have moral implications (Martinho et al., 2021a). For example, algorithmic advisory systems are being developed that aim to assist medical professionals in critical treatment decisions (Meier et al., 2022), or autonomous vehicles have to decide which traffic participant to target when an accident cannot be avoided (Bonnemains et al., 2018). Thus, such AI systems are not understood as neutral instruments but rather as ‘artificial moral agents’ (AMAs) that are, to a certain degree, capable of making and executing ethical decisions (Nyholm, 2023). Considering that technologies become ever more autonomous in their decision-making processes, scholars believe the conscious act of computing ethics is long overdue (Portmann & D’Onofrio, 2022) and stress the need for machines themselves to be able to behave in an ethically correct manner (Bringsjord et al., 2006). Also, from a consumer perspective, designing technologies ‘ethically’ is very important and increases their trust and satisfaction with the product (Morley et al., 2021). However, theoretical debates about machine ethics are no longer sufficient solutions (Segun, 2021). Similarly, high-level principles meant to guide (programmers in) the design process of pertinent technologies are limited in their practical applicability, which impedes effective implementation (Conitzer et al., 2017; Floridi, 2023). Thus, scholars in the field of AI ethics should focus on deriving “practical approaches and counteract a gap between formulated principles and their practical implementation” (Häußermann & Lütge, 2022; p.346). More specifically, the more pragmatic approach called ‘computational ethics’ is suggested, which aims to translate ethical principles into computer codes to ensure that corresponding AI systems are compliant with an ethical theory and functionally dynamic to operate in the real world (Segun, 2021).

Despite its centrality in the endeavor to move away from rather theoretical discussions towards actualizing ethical autonomous systems, computational ethics has received little attention in academic literature so far (Segun, 2021), with existing work mainly focusing on its technical implementation or related technical complexities (e.g., Todorovski, 2023). However, since “the greater challenge in the field of AI ethics is a philosophical rather than an engineering or technological one” (Tajalli, 2021; p.448), analyzing the ethical and societal implications of this method and the resulting AMAs should come to the fore in the inquiry. In accordance with this, other studies have highlighted ethical issues and ramifications related to (not) relying on normative theories (Geisslinger et al., 2023) or norms when building “ethically-sound robots” (Coggins & Steinert, 2023), thereby critically assessing specific benchmarks that could be

adopted when formalizing ethical principles within AI systems. However, many open questions concerning imminent (societal) consequences (Awad et al., 2022), as well as the legitimacy or “truth” of the overarching method of computational ethics itself, still prevail (Johnson & Verdicchio, 2023). Similarly, current debates on AMAs are characterized by controversy concerning the fundamental questions of whether these systems should even be developed and, if so, how to develop them (Martinho et al., 2021b). As the responsible development of AMAs coincides with the underlying method of computational ethics, this study adopts a bird’s eye perspective and explores the following research questions:

1. What are arguments for and against the use and development of artificial moral agents?
2. What are arguments for and against the method of computational ethics?
3. How can/should computational ethics and artificial moral agents be implemented?

These questions are answered by drawing on semi-structured interviews with twelve experts from the fields of philosophy, AI, and cognitive sciences. Given the infancy and division associated with the emerging phenomena of computational ethics and AMAs, qualitative expert interviews are a suitable research method to gather knowledge in an explorative manner (Bogner et al., 2014).

2.2 Theoretical background

2.2.1 (Designing) ethical decision-making of AI systems

That technologies (and their underlying algorithms) make decisions that carry ethical dimensions and hold ethical implications is not new (Mittelstadt et al., 2016; Moor, 1995). However, this phenomenon achieves ever more significance when looking at the development of emerging technologies: Today, some AI systems are now expected to make or suggest ethical decisions on their own completely. Thus, scholars argue that the existence of AMAs has become a reality (Nallur et al., 2023). For example, autonomous vehicles need to distribute risks between road users in traffic or algorithms in the criminal justice system to adjudicate an individual’s sentence (Awad & Levine, 2020; Geisslinger et al., 2021). “Designing such technology implies to a certain degree also to engineer human life” (Alt et al., 2021; p.9). Since it can be expected that such intelligent machines assume an increasingly prominent role throughout life spheres and industries, it needs to be ensured that these systems behave in ‘ethically correct’ (Bringsjord et al., 2006) and ‘fair’ manners (Awad & Levine, 2020). To do so, different approaches have been introduced in the past that can broadly be separated into ‘ethics in design’ and ‘ethics by design’. The former (i.e., ‘*ethics in design*’) refers to activities conducted by the development team during the design and evaluation process of technology, which aim at ensuring “social, legal and ethical

acceptability of the system” (Dignum, 2019; p.6). In this regard, expert groups and professional associations have formulated guidelines and ethical principles, such as fairness or explainability that programmers should adhere to when designing AI systems (e.g., Floridi et al., 2018; IEEE, 2019). Approaches such as value-sensitive design (Friedman & Hendry, 2019) or value-based engineering (Spiekermann & Winkler, 2022) put forward more specific and systematic methodologies to integrate ethics (e.g., in the form of design features) throughout the full life cycle of a technology. ‘*Ethics by design*’ corresponds to shaping the behavior of technologies themselves (Dignum, 2019), which entails installing ethical parameters directly into AI systems that inform their decision-making processes (Salo-Pöntinen, 2021). These technologies are then capable of engaging in some kind of ethical reasoning (Bonnemains et al., 2018) and ultimately become ethically aligned systems (Segun, 2021). The underlying method and scholarly work dealing with the codification of ethics has been coined as ‘computational ethics’.

2.2.2. Computational ethics

Computational ethics “seek[s] ways to translate abstract moral principles into computer codes [...] [and to] develop ethically grounded algorithms” (Segun, 2021; pp.20-22) and, thereby, “engineer ethical AI systems” and maybe even generate a better understanding of human moral decisions (Awad et al., 2022; p.389). In terms of AI systems, the ultimate goal is to ensure that they follow (sets of) ethical principles in their decision-making process so that they engage in ethical behavior (Nath & Sahu, 2020). When formalizing ethics in algorithmic terms, past scholars have mentioned that these approaches need to be *informed by human values* and grounded in ethical theory, which emphasizes the need to draw on a mix of descriptive ethics and normative ethics (Jacobs & Huldtgren, 2021). In this sense, descriptive ethics means “determining what people think is morally right or wrong, and [...] formalizing these views [...] in computational terms” (Awad et al., 2022; p.396). For example, previous investigations showed that study participants do not consider race as an acceptable attribute for prioritizing patients in kidney exchanges to be consulted in a corresponding allocation algorithm (Freedman et al., 2020). However, sometimes individuals’ preferences/values are (objectively) not valuable (van de Poel, 2020). To forgo committing to the naturalistic fallacy through ‘crowdsourcing ethics’, responsible oversight, and the *simultaneous consideration of normative theories* is helpful (Awad & Levine, 2020). Manifesting normative ethics means, for example, combining moral principles/rules, formal logic, or formalized conceptions of moral character with automated reasoning (Awad et al., 2022). Various philosophical theories, such as deontological theories, consequentialism, virtues ethics, or contractualism, have been proposed as the ethical benchmark of choice for technological systems (Tajalli, 2021). Technically, descriptive and normative ethics can be implemented through *different architectures, namely: top-down, bottom-up, or hybrid*

approaches (Bonnemais et al., 2018). Top-down implementing approaches “try to implement some specific normative theory of ethics into autonomous agents [e.g., in the form of rules] so as to ensure that an agent acts in accordance with the principles of this theory” (Dyoub et al., 2020; p.7). Bottom-up approaches rather correspond to the method of machine learning in that, amongst others, “machines are expected to learn how to render ethical decisions through observation of human behavior in actual situations, without being taught any formal rules or being equipped with any particular moral philosophy” (Etzioni & Etzioni, 2017; pp.406-407). Hybrid approaches are a combination of the two methods by, for example, supplementing the top-down deployment of rules with bottom-up contextual observations and learnings (Woodgate & Ajmeri, 2022).

Despite these existing technical approaches and the (inevitable) emergence of AMAs, the true compatibility of morality and computing remains challenged (Sinnott-Armstrong & Skorburg, 2021), leaving some barriers to successfully implementing computational ethics. In particular, it is questionable how exactly we would program ethical decision-making into machines and whether “ethics [is] the sort of thing that is amendable to programming” (Moor, 1995; p.1), amongst others, due to its abstract nature (Tolmeijer et al., 2020) and context-dependency (Awad et al., 2022). In addition to (or maybe precisely because of) these existing challenges, scholars have raised the question of whether we should pursue codifying ethics and the development of AMAs at all (Dignum, 2019). If we agree on pursuing computational ethics and the development of AMAs, there is a pressing need to create systematic ways of encoding ethical principles/rules into AI systems (Woodgate & Ajmeri, 2022) and develop “a coordinated effort by [...] politics, business and academia [...] to make AI a force for good” (Taddeo, 2019; p.190). This is precisely this study’s subject of investigation, namely to elaborate on the (ethical) reasons for and against computational ethics and resulting technologies, as well as how computational ethics and AMAs could or should be implemented in practice.

2.3 Methods

2.3.1 Participants

Due to the niche area and the explorative nature of the investigation (i.e., studying the emerging phenomena of artificial moral agents and computational ethics), we adopted the purposive sampling technique by conducting in-depth interviews with key experts (Reinecke et al., 2016). Key experts in this regard are individuals knowledgeable in the phenomenon of interest (Moser & Korstjens, 2018), who, amongst others, can be identified by the number of articles they have published in the respective research field (Guy et al., 2021). Thus, corresponding experts were searched via Google Scholar by consulting keywords such as “computational ethics”, “ethical engineering”, “ethical computing”, or related topics. The hits were automatically sorted

by relevance, and authors publishing about computational ethics, especially with a focus on its ethical ramifications, were considered as eligible for inclusion as interviewees. These individuals were contacted with a standard contact email introducing the study and research team and asking for their participation in an expert interview. In addition to this initial sampling process, we utilized the snowball technique by asking the interviewees for further contacts they could recommend (Goldstein, 2002). Overall, to derive a balanced sample in the sense of representing multiple perspectives on the topic, experts from different professional backgrounds (e.g., philosophy, computer science, psychology) were included in the sample as well as interviewees were asked to make referrals to experts who have differing (e.g., drastically critical) opinions about the subject matter. Through this additional snowball sampling technique, we aimed to prevent generating a potentially biased sample, which included experts that are particularly devoted to the subject matter (e.g., manifested in their research focus listed on Google Scholar). In total, we interviewed twelve experts from different research disciplines (for an overview, see Table 2).

ID	Professional role	Research field/discipline	Length of interview (min)
E1	Researcher	Digital ethics	43
E2	Professor	Ethics & Technology	49
E3	Assistant professor	Ethics	51
E4	Doctoral researcher	Cognitive Sciences	63
E5	Researcher	AI ethics	46
E6	Professor	Philosophy	70
E7	Assistant professor	Philosophy	Written response
E8	Assistant professor	AI Ethics, Computational Ethics	38
E9	Professor, lab director	AI, foundations of AI	63
E10	Postdoctoral scholar	AI, data science, ethics	42
E11	Researcher	Human-computer interaction	41
E12	Associate professor	Philosophy of Technology	41

Table 2: Overview of interviewed experts

2.3.2 Data collection

The recruitment and interviewing period extended from November 2022 to February 2023. The interviews were semi-structured, which means an interview guide served as the basis for all interviews while simultaneously flexibility was allowed to adapt to the conversation and the perspective/professional background of the particular interviewee. The interview guide consisted

of questions concerning the expert's personal experience with computational ethics, reasons for and against it, as well as its implications and implementation (e.g., *What are reasons for formalizing ethical principles within algorithms/technologies?*, *What good consequences would follow?*, *How can we implement computational ethics?*). Small adaptations to the interview guide were introduced between interviews to gain more information on particular topics of interest, while less fruitful or saturated avenues of questions were cut. The final version of the interview guide can be viewed in Appendix A of Essay I – Interview guide. At least one day before the interview, all interviewees received study leaflets with in-depth information about the study (i.e., the declaration of consent, a preamble text defining the key concepts essential to the interview, as well as the interview guide). A Zoom call was set up with eleven of the participants. One interviewee (E7) responded via email by filling out the PDF document of the interview guide. Written and informed consent to participate in this study was obtained from all interviewees. To build trust and to allow a higher probability of uncovering sensitive data and opinions, we ensured all interviewees' anonymity – unless preferred otherwise – in the data analysis and publication process. At the beginning of all interviews, the interviewer again quickly introduced the topic of the interview and the data processing and gave the respondents time to ask any question concerning the study before the audio recording started. The first author conducted all interviews to maintain consistency. Recruitment of additional experts ended when saturation was consistently reached, meaning when subsequent interviews no longer brought up new themes in a substantial manner ($\leq 5\%$) (Guest et al., 2020).

2.3.3 Data analysis

All interviews were audio-recorded and transcribed verbatim into a computer file (547 minutes; 105 pages). The text was then manually coded using the coding software MaxQDA. Given that much of the prior research on computational ethics focused on its technical implementation and not on its ethical admissibility or societal consequences, we had a limited basis for assigning a priori categorizations to the data. In such cases, inductive methodologies are most appropriate, allowing researchers to generate novel theories from complex phenomena (Gehman et al., 2018). After all, “it is impossible to know prior to the investigation what salient problems or what relevant concepts will be derived from this set of data” (Corbin & Strauss, 1994; p.36). Therefore, we drew on inductive coding methodologies (Corbin & Strauss, 1994; Gioia et al., 2012) as opposed to a deductive coding methodology that draws on a template or an existing codebook as a means of organizing and interpreting interview passages (Fereday & Muir-Cochrane, 2006). Without predefined categories, we developed theoretical categories and identified themes as they emerged during data collection (i.e., directly from the interview transcripts). As a first step, we engaged in open coding and unrestricted labeling of data by reading

all transcripts line-by-line in their entirety and by encoding single text segments and passages as first-order concepts. As the coding process progressed, we iteratively reflected earlier developed codes with new data. In line with Gioia et al. (2012), we then organized 1st-order codes into 2nd-order themes/type of argument (i.e., with a slightly higher level of abstraction) from which we distilled aggregated dimensions (i.e., the overarching question that is addressed with the text passages). In addition to our manual coding, we utilized the automatic code generation function that is offered within the MaxQDA software⁹ (i.e., AI Assist). After the exclusion of incorrect, fuzzy, or irrelevant codes (e.g., codes that are too broad and immanent/specified within other code suggestions or codes that focus on the interviewee's manner of speaking), AI Assist resulted in 482 additional codes (see more in 2.3.4). Furthermore, we conducted frequency analyses of the 1st-order concepts (i.e., percentage of experts that indicated a particular concept), which can help to substantiate the weight or importance of a certain concept (Mayering, 2014) and complement the purely qualitative perspective (Pole, 2007) (see Appendix B of Essay I – Frequency analyses of the experts' arguments & recommendations). As a last step, in line with Gioia et al. (2012), we synthesized our coding structures into a (dynamic) data structure. The coding structures and frequencies are illustrated in Figure 6, Figure 7, and Figure 8, while the dynamic data structure is displayed in Figure 9.

2.3.4 Data validity

Before data collection, previous versions of the interview guide were discussed with experts in the field of qualitative research (e.g., they were asked to provide general feedback on the interview guide, such as its structure, or to rephrase the stated questions to determine whether our intention of asking a particular questions aligns with the understanding of respondents) and pretested with representatives from our target sample (i.e., researchers in the field of computational ethics). After the data collection and preliminary analysis by the first author, the developed codebook (consisting of code labels, their definitions, and corresponding interview passages) was reviewed by other research team members. In particular, the co-authors challenged and interrogated the coding results and interpretations of the first author and engaged in in-depth discussions to reach a common understanding for developing consensual interpretations and the ultimate data structure. Furthermore, we ensured higher levels of reliability and objectivity of our interpretations through triangulation with external researchers (Silverman, 2015) who were not involved in the research study (i.e., blind to our research question) by providing them with sections from the interview transcripts and asking them to map these to codes within the

⁹ MaxQDA AI Assist is a virtual research assistant that can create AI-generated summaries and analyses of various data formats, such as text documents. It is powered by OpenAI and specifically tailored for qualitative and mixed methods research purposes. (Retrieved from: <https://www.maxqda.com/>)

codebook. In addition, we validated our manually generated codes by consulting the automatically generated codes of AI Assist¹⁰. In particular, we read all automatically generated codes and their description and compared and mapped them to previously detected first-order concepts, which we subsequently extended where necessary. Moreover, once the first version of this article was drafted, we conducted interviewee checks by sending it to our interviewed experts and asking them for general feedback and to evaluate whether our results represent what they highlighted during their interview. Lastly, the findings of this investigation were presented and discussed with the science community at academic research seminars or international conferences such as the 2023 Forum on Philosophy, Engineering & Technology.

2.4 Findings

The expert interviews showed that when discussing the formalization of ethical principles within AI systems, two distinct (but related) questions and examinations need to be considered: the investigation of *whether artificial moral agents should be developed or utilized for ethical decision-making* at all (see 2.4.1) and if so, *whether computational ethics should be utilized to develop pertinent technology* (see 2.4.2). For both examinations, experts indicated supporting and opposing arguments that can be clustered into practical, societal, and epistemic reasons. Practical arguments address the technical feasibility and demand/need for AMAs or computational ethics. Societal arguments entail societal preconditions that endorse the use of AMAs and the societal implications of these systems and computational ethics. Epistemic arguments comprise the (reciprocal) impact between AMAs or computational ethics and human knowledge. In addition to arguments for and against AMAs and computational ethics, experts *recommended how to develop artificial moral agents and implement computational ethics* (see 2.4.3).

2.4.1 Artificial moral agents for (supporting) ethical decisions?

This section summarizes all arguments for and against developing/utilizing AMAs for ethical decision-making that experts indicated in our interviews (see Figure 6), thereby answering our first research question.

Supporting arguments

From a *practical perspective*, experts stated two supporting arguments for developing and using AMAs for ethical decision-making:

¹⁰ Based on machine learning and AI techniques, datasets can be precisely examined, ultimately “identifying patterns and insights that were once beyond reach” (Bryda & Costa, 2023; p.1). This is why the analytical validity in qualitative research can be fostered by using computer programs such as MAXQDA (AI Assist) for transcript and data analysis (Demir-Kaymak et al., 2024; Whitemore et al., 2001).

Technology's efficiency and precision in making and executing decisions. 42% of the experts mentioned the superiority of technology (compared to humans) by referring to systems such as autonomous weapons or autonomous vehicles. For example, autonomous weapons may make more accurate and faster predictions of a target's surroundings so that fewer casualties of innocent civilians will occur (E6). The ability to make quicker decisions than humans was also mentioned in the case of autonomous vehicles in traffic (E6, E11). In addition, the use of technology for decisions is connected to creating decisions that are more aligned with each other and are universally made:

"And you know, when you put things into code, there is this possibility of scaling up. This is one of the advantages that you can scale up decisions. So now you can do the same thing in multiple different places at the same time." (E8)

For autonomous vehicles operating in a fully connected traffic infrastructure, this would entail that the vehicles are "all thinking in the same way as opposed to people who have different competing priorities" as well as be in constant communication with each other, which may lead to higher levels of safety on the road (E11).

Corresponding technology is already in use. This type of argument was quite pragmatic, addressing the staleness of evaluating whether technological systems should be utilized for ethical decisions, as they are already being developed and used in practice anyways. When we asked one of the experts why they focused on decision-support systems in the field of healthcare in one of their studies, the expert answered:

"We picked kidney transplants, largely because number one: We knew that computers were already being used in kidney exchange programs." (E6)

Overall, mentioned examples of prevalent technology that is deployed for (assisting) ethical decisions span from the legal domain (i.e., systems that help make court judgments or set parole for individuals), the business context (i.e., systems that help make recruiting decisions), transportation (i.e., autonomous vehicles), healthcare (i.e., allocation tools for scarce medical resources or well-being apps) to the military domain (i.e., automated drones) (E3, E6, E7, E8, E10, E11, E12).

From a *societal perspective*, experts indicated the following two beneficial outcomes:

Calibration of flawed nature and decision-making of humans. 67% of the experts stated that humans by nature are "fallible beings" (E2, E6) and make many normative mistakes, even in day-to-day decisions, such as when determining what products to consume (E5). Engaging in poor decision-making can be attributed to human's "bounded ethicality"

and “hidden psychological forces”, which lead to individuals not following through with moral or rational behavior despite having strong moral intuitions of doing so (E5). Thus, one expert stressed the need to accept that AMAs can support humans in their decision-making:

“This idea that humans are always going to do it better than computers strikes me as just pride, arrogance, and so we have to admit that we make mistakes and see if we can use computers to help us make fewer mistakes.” (E6)

The emphasis here is on technologies assisting (not taking over completely) individuals' decision-making so that they are directed towards reflecting on their own actions and identifying their own potential blind spots/biases, thereby reducing human prejudices (E3, E5, E6). In this regard, one expert pointed to the example of medical treatments:

“Studies have suggested that doctors often give less painkiller to people with dark skin, and you tell the doctors they're doing that, and they go: No, I would never do that. That's wrong. I don't want to do that. But they are.” (E6)

Another mentioned example related to the inconsistency between doctors in making medical decisions due to their differing levels of or responsiveness to empathy towards particular patients, for example, those who “cried a little bit more” (E8). Thus, the overall goal of technological decision-support systems would be to help humans make less biased, more reflected, and coherent decisions (E8, E10) that are “in line with their own values not imposing a different set of values” and are results of sufficient information and impartiality (E6).

Liberation of users' capacity to engage in other decisions and tasks. 17% of the experts addressed how AMAs could free humans from particular decisions and, thereby, create capacity for them to engage in different decisions and tasks. For example, on a general level, humans would have more time available for “more interesting or more creative” tasks (E8), or technological applications in clinics would reduce the workload of nurses so that they can engage in more interpersonal connections with patients (E11).

Improvement of societal well-being. Similarly, 17% of the experts indicated that – amongst others, due to the superiority (e.g., in terms of efficiency) of technologies compared to humans – the implementation of AMAs could lead to higher levels of safety in traffic with autonomous vehicles, enhanced quality of care in hospitals with care robots (E11) or less bias and inequalities overall (E6).

From an *epistemic perspective*, two arguments were brought forward by experts that speak for AMAs:

Achieving confidence in human decisions via additional justification/crosscheck. In this regard, 17% of the experts stated that AMAs could serve humans as crosschecks or justifications for their decisions, thereby generating confidence in their decisions. For example, when a doctor has to make a distribution decision:

“Let the computer do it. And when you agree: Hey! You know, now you're more confident and justified in being more confident. When you disagree: Go ask somebody else as a consultant to make sure that you're doing the right thing. You know that kind of use of the computer, it seems to me, is not going to create problems.” (E6)

Even in the long run, individuals will not lose the ability or confidence to make ethical decisions when supported by technology because “these machines are going to be doing it in very limited contexts”, according to this expert (E6).

Transparency over how a decision was reached and can be adjusted. 17% of the experts stated that compared to humans, technologies offer higher levels of transparency in their decision-making process and the means of readjustments:

“Of course, they [machines] have their biases and all of these things. But some people argue that fixing machine bias is easier than fixing human bias.” (E8)

That is because, generally, algorithms work on a “set of if-and-else” functions (E8) or based on underlying data that can easily be inspected and changed (E6). In turn, this may lead to higher acceptance levels and “less resentment [and] skepticism” concerning a particular decision because it is the result of a technological system instead of a human (E6).

Opposing arguments

From a *practical perspective*, experts indicated only one argument against developing and utilizing AMAs for ethical decision-making:

Inflexibility and lacking timeliness of technology. According to 17% of the experts, there is a discrepancy between regularly changing contexts and morals and technologies' inability to flexibly and in due time respond to and adopt these: “Once I've trained an AI system, it's pretty inflexible” (E5). Over time, this will lead to a misfit between the previously embedded values or design requirements and prevalent morals, so algorithms may need to be retraced (E12).

From a *societal perspective*, there are many barriers or repercussions that need to be critically considered when adopting AMAs, according to the interviewed experts:

Overreliance/trust of humans in technologies' decision-making and suggestions.

Experts stressed that users “come to rely on the system in a way that we cannot rely on them” (E2), amongst others, by adopting provided suggestions in an unfiltered manner:

“Us as people, we have the orientation to go into an automatic mode when we interact with intelligent machines. And this automatic mode makes it for us very hard to acknowledge when, for example, the decision-support system provides us with fallible solutions, so we are not capable of detecting them.” (E4)

This is particularly dangerous when suggestions are “persuasive but unsound” (E9) or “way off base”, but individuals still go along with them (E6). Relying too much on such systems when making decisions may, on the one hand, result from convincing technological devices giving “spurious argumentation for actions that should be taken” (E9) or due to human's loss of making independent ethical decisions over time. The latter was accentuated by the following example: “If you get so used to automated vehicles that they just drive you to work, you might get worse at driving” (E6). As another reason why individuals rely too much on decision-support systems, one expert highlighted individuals' lack of motivation or “moral laziness” in that delegating decisions to technology represents a comfortable solution (E12).

Technological determinism via autonomous development of (im)moral behavior.

50% of the experts mentioned how technological systems themselves could develop their own (im)moral behavior that humans cannot anticipate or control. This is particularly the case for systems operating on machine learning techniques that retrospectively (i.e., after implementation) find or create patterns that differ from previously designed/embedded ethical principles:

“AI systems [...] can be adaptive systems that are self-learning. And, in principle, the systems could also disembed the value in the way they learn. I mean, you can design a system that is safe or accountable. But if the algorithm starts to work differently, maybe it is no longer safe. [...] You cannot predict how these systems can change themselves.” (E2)

Ultimately, these machines could make up and enforce their own principles and decisions (E6). For example, autonomous vehicles may learn “making a lot of stops, which makes it maybe not good for a human experience” (E8). Phenomena like these were perceived as a “threat to democracy and the core values of our modern society” (E10), especially when these decisions are “immoral or unacceptable” (E2). Nevertheless, when it comes to corresponding technologies and the development of these technologies, “we seem to be ourselves much more paternalistic” in the sense of allowing others to decide for us (E12).

Potential generation of (physical) harm to humans. Since technologies – just like humans – also sometimes make mistakes, it is inevitable that individuals will get hurt when AMAs execute decisions. Potential harm spans from physical injuries to violations of human rights (E6), for example, by discriminating against individuals based on their background (E12). This is why the use of AMAs should be restricted in “crazy, [...] dangerous, [...] really destructive things” (E9).

Creation of moral patiency for technology. One expert highlighted that providing ethical agency to technological systems may simultaneously implicate their claim to moral patiency (i.e., having rights):

“If you endow a machine or a computer with moral reasoning, you are making it a moral agent, and therefore, you risk making it a moral patient. And so a moral patient is something that is deserving of moral consideration. So if we have an autonomous car that is making an ethical decision, then expecting it to sacrifice itself is unreasonable.” (E11)

While the expert emphasizes that an autonomous vehicle is not a “biological thing”, thinking of moral patiency for technological artifacts is still worth pursuing as thought experiments and may be required in the future (E11).

From an *epistemic perspective*, experts stated that using AMAs may be associated with the following two destructive consequences for human knowledge:

Spreading the misconception that technology can be ethical and autonomous. Only 33% of the experts highlighted that offering AMAs for ethical decisions may create the delusion/misunderstanding that technologies (similar to living beings) can be considered ethical and can act autonomously, which, however, “does not make any sense at all” (E1). Namely, experts warn their capabilities should not be equated with these technologies themselves being ethical, as this, amongst others, entails “being capable of taking responsibility for its actions, experiencing moral emotions, being worthy of praise or blame” (E7). These capabilities and morality overall “is intrinsically human” and thus not applicable to technological artifacts. Instead, technologies should be perceived “as having low autonomy and still low ethical sensitivity” (E10). Another expert highlighted being wary about the discrepancy between what such technologies are “promising [...] and what they are actually delivering” (E3). One expert referred to another scholar who even said, “there should be a ban on research in machine ethics because this will lead companies to think we can build ethical machines” (E2).

Creating intransparency through accountability/responsibility gaps. These indicated arguments addressed the ambiguity and diffusion of responsibility and accountability

between human users and technology with the adoption of AMAs. One expert pointed out this concern by directly raising the question: “Who is going to be the ultimate person accountable for moral decisions that machines are making?” (E11). Compared to when an individual independently makes a decision, the process of identifying the responsible and accountable party is less straightforward when a decision-support tool is used:

“For humans, we have clear processes of holding them accountable if they make a mistake. But we cannot hold a machine accountable. We can maybe hold the producer of the algorithm accountable. [...] Some of our practices for holding accountable or blaming if something goes wrong don't work for machines.” (E2)

What can happen is that eventually, “responsibility gaps” will occur (E7), especially because developing technological companies will aim to prevent accountability from falling back on them (E11).

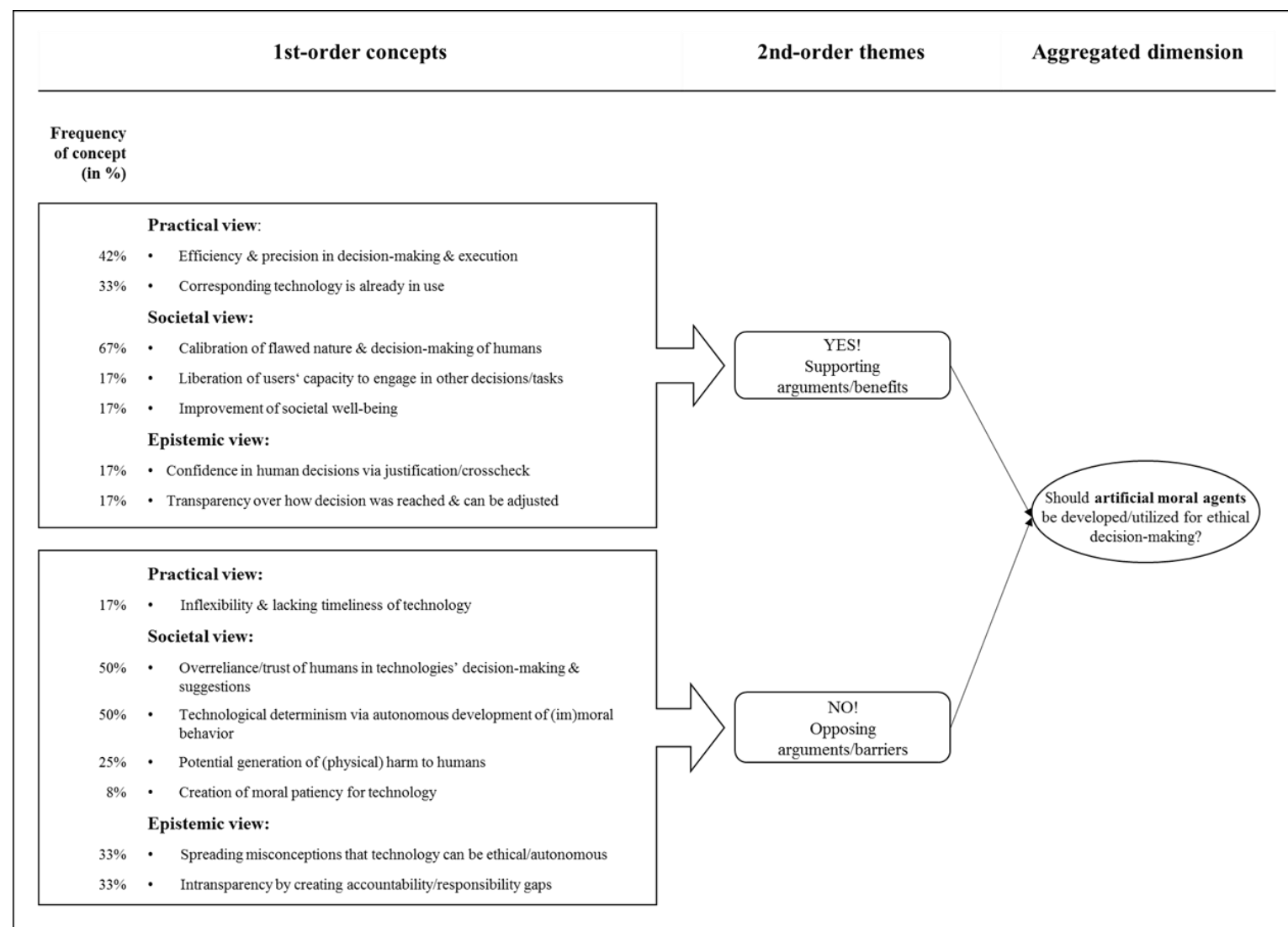


Figure 6: Experts' indicated arguments for and against developing/utilizing artificial moral agents for ethical decision-making (ordered by perspective and frequency)

2.4.2 Computational ethics for developing artificial moral agents?

This section summarizes all arguments for and against utilizing the method of computational ethics to develop pertinent technological systems that experts indicated in our interviews (see Figure 7), thereby answering our second research question.

Supporting arguments

From a *practical perspective*, experts indicted the following reasons for why it is feasible and important to formalize ethical principles within algorithms/technology:

Practitioners acknowledge the need, effect, and responsibility to design for values.

A few experts agreed that companies are aware that what they are programming and their products themselves have social and ethical implications, so they are “open to the idea of designing for values, etc.” (E3), which makes them also more receptive to and call for solutions such as computational ethics (E9, E11). This new endeavor of computational ethics allows ethics to be turned into an applied philosophical branch:

“This is a fascinating time for the normative communities, for ethics in general. Ethics used to be very far removed [...] from practice, and it's fascinating to be in this moment in time because there's a clear need, and it is accepted by almost everyone that there is a need for ethics, and that is an opportunity.” (E10)

To meet this need, ethical training is offered at universities or at companies so that (future) engineers and programmers are knowledgeable of ethical considerations when developing corresponding technology, compared to “old school practitioners that [...] are not able to have a holistic view about society and how technology may have a pervasive impact” (E10).

Future regulation and societal pressures may require companies to design for values. 58% of the experts mentioned that practitioners might soon be obligated and pressured to consider the method of computational ethics. For example, media and policy discussions (e.g., on potential regulatory measures) (E3, E4), as well as societal attitudes towards the importance of ethics in technology (E1), have contributed to the momentum.

Current technologies are exposed to moral situations and need to respond. Experts highlighted the inevitability and necessity of implementing ethics within algorithms/technology since they are or will be placed in situations anyways that require them to make (ethical) decisions (E10). One expert argued:

“We're going to have to put it internal to the machines as they get more and more autonomous, more and more intelligent, and as they find themselves situated in ethically charged situations. We're gonna have to put this stuff into the machines. There's no getting

around it. [...] We can deceive ourselves and say that we don't have to, but we're gonna have to.” (E9)

For example, when technologies make “important decisions or judgments about people’s lives (e.g., hiring, admissions, parole decisions)” they would need to comply with requirements and operationalized aspects of fairness or justice (E7). While the actual adoption of such technological systems beyond test runs and without human oversight is still limited, they are being built and prepared to be placed in moral dilemma situations in the future and then make independent decisions (E8).

From a *societal perspective*, three supporting arguments for computational ethics were stated by experts during the interviews:

Forgoing technological determinism by proactively controlling embedded ethical principles. While the emergence of technological determinism was associated with such technological systems, 75% of the experts mentioned that computational ethics could be viewed as one way to avoid this phenomenon by allowing humans to determine the values that underlie a technology’s functionality themselves. In this sense, individuals could regain control over the logic and outcomes of technological systems:

“In effect, what computational ethics does is it puts the break on the AI doing something. It stops it from doing something that people would think is unethical.” (E6)

Thus, instead of “basically letting it happen as a side effect” (E8), through computational ethics, individuals could shape the functioning of these systems in a “premeditated fashion” (E9) “to ensure that they don’t develop in undesirable ways” (E2). One expert, however, put the human level of control into perspective by distinguishing between ‘principles’ and ‘rules’ and thereby stressing that picking a value or morally relevant features, which a technological system should take into consideration when making/suggesting a decision, does not necessarily “determine the outcome” (E6).

Consideration of morally relevant factors leading to ‘fair’ societal outcomes. Overall, “it would be better if the systems were responsive to morally relevant aspects of the world” (E7) or of a user’s environment (E9). Correspondingly, 50% of the experts mentioned that computational ethics would allow the regard of morally relevant factors within a decision situation so that the technology arrives at a more fair societal decision. As an example of such a morally relevant factor, one expert named the presence/close location of a hospital to the target (i.e., a terrorist group) an autonomous drone is supposed to strike: If such factors would be neglected, “you’re gonna kill a lot more civilians” (E6). Similarly, by referring to autonomous vehicles, the experts said:

"Mercedes was famous for it at one point when one of its spokespeople said: We're going to build our cars so as to minimize the chances of death and injury to the people IN the car. Whoa! Like that's not building an ethical principle in at all. That's just going: Well, they bought the car, so we're going to give them what they want. And that's going to be a real problem for pedestrians because people in the car have airbags, and pedestrians don't have airbags. [...] If we build in ethical principles, we'll have fewer people killed in accidents that should not have been killed." (E6)

With this example, the expert addressed how computational ethics may promote fairness in the distribution of harm between parties affected by the use of a particular technology. Another expert highlighted the importance of considering computational ethics by remembering an automated system that neglected specific factors in its decision-making process and, subsequently, distributed childcare beneficiaries unequally (E4). Thus, to achieve fair outcomes and distributions of benefits (and drawbacks), it is crucial to incorporate values "since the beginning of the design phase to ensure that the technologies [...] have socially good outcomes" (E10).

Creation of public acceptance of artificial moral agents. 25% of the experts pointed out that formalizing ethical principles within algorithms/technology may lead to public acceptance and a higher willingness to adopt pertinent technologies. One expert stated:

"If we consider it [computational ethics] as part of designing desirable technologies, it helps us to take advantage of the potential of AI technologies and automated systems as it helps to make them more acceptable for the larger public." (E4)

By contrast, if a technology produced a bad outcome (i.e., an accident) and it turned out that ethical considerations or morally relevant factors were not integrated into its decision-making process, it would be "a disaster from a public relations perspective, and nobody is going to trust you anymore" (E6).

From an *epistemic perspective*, engaging in the endeavor of formalizing ethical principles within algorithms/technology can have positive impacts on human's state of information in the following ways:

Education on ethics/morality in general. In addition to learning about computational ethics itself, 42% of the experts stated that through its inquiry, also knowledge about ethics and morality can be generated. One expert highlighted:

“It is important to formalize and quantify ethics because this will help us actually think more clearly and force us to make decisions about what kind of ethics, what kind of moral decisions we actually make in our daily life.” (E8)

For example, when an AMA is trained with a bottom-up approach, based on its outcomes or suggestions generated, “this system could give us an insight into human morality” as it practically mirrors societal preferences (E5). Similarly, if human normative cognition were formalized within computer systems, “successful simulations [...] may point us to interesting hypotheses about how parts of human cognition might work” (E7). Thus, particular technical implementations of computational ethics may contribute to additional knowledge generation in descriptive ethics (i.e., human morality). Furthermore, computational ethics adds another layer to the investigation of ethics overall:

“Researching or exploring and theorizing more about ethics in different contexts is a good human endeavor to try and understand ourselves better, to understand their own ethics when they're thinking about applying somewhere else.” (E11)

Applying ethics and “putting it into machines and code will probably [...] force us to make a decision” of what is right and wrong, thereby creating more clarity on how to respond to particular ethical issues (E8). Thus, “for philosophy [computational ethics] would also be interesting” as we may come to understand better what moral judgment really is or should be (E2).

Learning about the feasibility and implications of computational ethics. A few experts (33%) highlighted that “computational ethics is a very, very interesting endeavor” (E4), especially as a “philosophical or scientific exercise” (E2). But not only for the research community, corresponding investigations would lead to significant findings: Experts stated that the inquiry of computational ethics is necessary to study if/how it is possible at all and to determine what consequences would follow. One expert indicated this by saying:

“Well, it's interesting from a scientific point of view, that is to say: Well, we try to do it and see if it is interesting or not. So we have to [...] try and implement that to see whether it is worth doing it or not.” (E1)

Thus, without exploring it, we will not find out “whether ethics could really be implemented in the algorithm” in the first place. Instead, we can only “really learn a lot by doing and then seeing” (E3). One expert referred to their own previous investigation in which they “tried to model ethical frameworks in a logical way and see whether this was meaningful or not” by identifying what kind of knowledge was represented in machine learning and where corresponding limits occurred (E1). In addition, such investigations could bring insights into

what we “can and cannot understand” or into the reasons for technological failures, for example, in which situations or “where it [building moral principles into artificial agents] goes wrong” (E2).

Opposing arguments

From a *practical perspective*, the interviewed experts brought up many arguments against computational ethics and corresponding barriers:

Limited ability to code ethics in its entirety due to its context-dependence and complexity. 83% of the experts argued that the totality of what makes up ethics could never be grasped or coded since ethics is complex and varies with context. For example, one expert stressed the difficulty for a machine to simulate human morality and our tendency to underestimate “the complexity of human moral and normative cognition” (E12). Encoding and ultimately sticking to principles cannot be equated with ethical behavior in that “moral judgment is much more than [...] applying moral principles” (E2). Instead, context plays a significant role in ethical decision-making (E3, E4, E12), which “falls short” when trying to operationalize ethics (E2) because it is impossible for humans (e.g., programmers) “to fully comprehend every single piece of context” (E11). Although “taking into account the real situation in real-time” would be necessary, whatever is actually formalized and integrated into technology can rather be considered as “simplifications [...] of much more complex concepts” (E1). Thus, as soon as “we tokenize the data, we lose declarative information that’s never going to be retrieved” (E9). In line with this, one expert raised the open question:

“You have to leave certain things out, or you have to focus on a specific aspect and therefore not pay as much attention to other aspects. [...] How can we account for the messiness of reality and all the details that make stuff super complex and give rise to ethical dilemmas?” (E3)

As one potential solution, one expert emphasized that what is considered the ethical thing to do should not be established across different contexts:

“We cannot generalize ethics. Ethics is applied to one situation at a given time, and it has to be revised constantly. So you cannot say: Well, in this situation, this is what should be done. And in almost the same situation: Well, this is what should be... No, this is not how it works!” (E1)

Low demand of/resistance by companies to implement computational ethics. In contrast to practitioners’ acknowledgment of the importance to design for values, 67% of the experts also added the limited demand or resistance of companies to actually formalize

ethical principles within their algorithms/technology. One expert confirmed this tension by remembering their own experiences with practitioners:

“They were really quite resistant to the idea of an ethics-by-design approach that follows certain steps and is based on certain moral principles and guidelines and values. So, on the one hand, they did want some ethical input, etc. But as soon as that ethical input was somewhat formalized into a framework or an approach, it was really like: ‘No, no, no, no. We want to be free [...] and creative in our process’. [...] The idea of really formalizing ethics: I am not quite sure whether they would be very open to that. [...] As soon as they had the idea of, ‘I have to follow certain steps or certain guidelines or principles’, they became very resistant.” (E3)

Another expert shared a similar experience by saying that during meetings with industry practitioners, they considered ethics an important topic, but “discussions rarely got to the point of computational forms of ethics” (E4). For example, when looking at the manufacturing of autonomous vehicles, no single company currently works on ethical decision mechanisms for cars (E5). Instead of directly implementing ethical principles into technology, which is contemplated with “fear” and “push back” (E9), “[t]he ethical issues the companies are dealing with are simpler, so they are dealing with privacy, explainability, fairness” (E5). Thus, the focus of the industry is on ‘ethics in design’ in that ethical principles are rather considered for humans as “guides to better engineer AI”, while “we’re talking very little still about putting these sensitivities or sensibilities into machines themselves”, which was perceived as a problem (E9). Another reason why the industry refrains from engaging in computational ethics is that machines only need to comply with the law and regulations (such as the Highway Code when it comes to autonomous vehicles) but not with ethical guidelines (E1, E10). Thus, companies may only be more receptive and accepting of computational ethics once “the ISO put out an ethical machines code that said [...] all machines have to follow these certain ethical principles” (E11).

Limited sophistication/reliability of technology and computational ethics approaches. 67% of the experts stated that current technology or computational ethics approaches need to be more mature to successfully execute ethical decision-making. For example, sensing techniques of autonomous vehicles may not be good enough yet to grasp entire surroundings and determine the best vehicle behavior: To emphasize this, one expert remembered “this case from this Google car that crashed because they simply couldn’t distinguish a tree from a pedestrian” (E2). Therefore, implementing ethical principles and morally relevant factors that the technology should take into consideration when making decisions would not have any effect in practice. Technologies that “rank very high [...] in

ethical sensitivity” can rather be considered “part of science fiction today” (E10). Furthermore, as one reason why ethical considerations can be disregarded within technologies’ algorithms, this expert mentioned:

“[I]mplement ethical principles into machines...I don't think this is necessary so far because the systems that are actually developed are only applied in so narrow contexts that there is no real requirement for the machine to have a sophisticated understanding of ethics.” (E5)

According to this expert, as long as today’s technologies only function in restricted areas, they will not need to operate on computational ethics. In addition, experts argued that technologies that have ethical principles implemented in them are not yet applied in practice and “they are still in the labs” (E1). This may be because the computational ethics approaches themselves are “unsophisticated” or not “reliable enough” and, thus, “really far away from something that you want to put out into society” beyond experimental settings (E2).

Existence of a plethora of (abstract) ethical guidelines and theories. 50% of the experts addressed the multitude of ethical guidelines that must be turned from abstract theoretical frameworks into formal logic. One expert hypothesized that this multitude of existing guidelines might have the consequence that technology companies “can't remember what sort of ethics they meant to be following when creating machines” (E11). Another expert explained the past development of ethical guidelines in the following way:

“There was a proliferation of principles, guidelines that were designed [...], even from industry. And I think that the major challenge today is not about not having resources as having too much resources. So, there's a plethora of guidelines and principles that exist out there. But then we need to kind of start trimming them down and start operationalizing them so that they can be used. So, I would say that that would be the major challenge nowadays.” (E10)

Thus, this expert added that the research community should – instead of focusing on “too general” work – start developing concrete “deliverables so that this is not a missed opportunity for them” (E10). Another expert highlighted the challenge of truly applying theories in practice (i.e., within technology). For example, it is challenging to implement ethical theories, especially when these are from different “families of ethical theories” such as consequentialism or virtue ethics, as, in practice, they will turn out to be “pairwise inconsistent” (E9).

Missing expertise and network of practitioners to formalize ethics. 42% of the experts addressed the limited skills of practitioners and missing collaborations to work on and

implement computational ethics. One expert highlighted this scarcity by pressing the following question:

"Where is your labor to translate that into working AI so the systems are abiding by it? And the answer to that is: Well, those people are hard to come by." (E9)

The expert continued and said that the talent pool capable of formalizing ethics or principles such as fairness or non-discrimination is "in very short supply", so society is unable to "do what we need to do in the area of machine ethics" (E9). Similarly, in the academic community, expert committees in the field of computational ethics are still underdeveloped:

"One challenge to doing research on the topic of CE is that there is not currently a network (to my knowledge) or regular conference connecting the researchers from different disciplines who are interested in or working on this topic." (E7)

Therefore, the experts stressed the need to "address this labor issue" (E9) by, for example, training the ethics community to be able to adopt such new methods (E10).

Technological solutionism: Primary focus on programmable principles and methods. 33% of the experts indicated that implementing computational ethics may lead to technological solutionism in that focus will be placed on only those ethical principles that can be formalized while neglecting other non-programmable ethical principles. As already discussed in the previous argument that was brought forward by the interviewed experts, "ethics in its totality is not lendable to computation" (E4). For one, this is because not everything can be put into numbers, as one expert elaborated by providing the following example:

"Think about the end of life in a hospital. Well, you cannot put that into numbers. This is a shared decision of medical doctors, family, etc. You have to put forward arguments. You do not decide on numbers." (E1)

This expert gave another example by referring to the principle of proportionality, which cannot be programmed in terms of "0 and 1" (E1). Thus, a threat of computational ethics would be to "neglect those ethics that cannot be implementable" or would be easy to do so, although these concepts are worth looking into (E8). In fact, one expert mentioned how this "bias" is already in effect with a focus on consequentialist theories:

"Everybody thinks of this particular framework [utilitarianism] because it's simple as you can put that into figures, so five, one, etc., and compare that. But [...] that's only one framework. What about deontologism, virtue ethics, the doctrine of double effects? We cannot reason only on the consequences and the numeral consequences of an action." (E1).

Overall, through engaging in computational ethics, “there is the risk of narrowing down what we mean with ethics, what we mean by implementing ethical ideas in technologies”, namely, through the means of ‘ethics by design’ (E4), while neglecting other approaches such as ‘ethics in design’. Another expert similarly perceived that focusing on the former path (i.e., computational ethics) alone is “scary” as it will signal to the developers: “We developed an ethically sound algorithm, so everything, all the outcomes will be fine now”, and no other design measures will need to be taken (E3).

Arising costs: Time-consuming & expensive activity. 25% of the experts addressed how formalizing ethical principles within algorithms/technology involves economic costs for companies. Similar to the establishment of ethics boards, practitioners could perceive computational ethics “as a nuisance, like something annoying” and as “an extra burden” (E3). Especially, “the more techie sort of people, the people that are actually involved in developing robots” emphasized that activities like these would create too much workload, as one expert recalled from their own interviews with practitioners (E11). Another expert highlighted that “ethical reasoning is very computationally expensive” because it needs extra time and because those individuals who are capable of doing so have to be paid “a ton of money”, which in turn makes the whole process “economically expensive” (E9).

From a *societal perspective*, the interviewed experts referred to the following opposing arguments against computational ethics:

Manifestation/reinforcement of human bias and preferences. 75% of the experts emphasized by adopting technologies that directly implement ethics into their functioning, human biases and preferences are internalized and reinforced (E3). These biases can spill over into the technologies in different ways and stem from various actors, depending on the underlying programming method. On the one hand, with bottom-up or machine learning approaches as a form of “crowdsourcing ethics”, biases from society may be implemented: What we then could end up with is a “mob rule in ethical decision-making” and “striking parallels in the limitations of the moral rationality in humans and machines” (E5). As a result, the technology would sometimes mimic and adopt decisions that are immoral (E2) or “less optimal” (E10). For instance, one expert referred to machine learning and explained its effect like this:

“You could base it on learning over data and massive amounts of data. [...] But the data inevitably is going to be ethically contaminated. It's going to be ethically contaminated because it's going to be based on the general space of human thought and behavior, and we have a lot of bad...a history of a lot of bad ethically bad behavior everywhere.” (E9)

As one example, the use of a chatbot was mentioned to which users could provide feedback, which resulted in the chatbot becoming racist (E5). Another expert highlighted a similar issue of discriminatory decision-making in relation to autonomous vehicles:

"If you're a person who [...], let's say, subconsciously disadvantages cyclists when they drive. And [...] you're just one person, maybe there are a few people like you. But if you implement that in code, then you have, like a hundreds of thousands of cars that do the same thing, and then this disadvantaging becomes more pronounced." (E8)

On the other hand, also with top-down approaches, biases may be implemented: these are the biases of the technology's programmers (E8) or "people who are, so to say, ethical experts" when they filter and decide which (training) data or labels to include and exclude (E5). Overall, what can emerge when incorporating ethical decision-making into technology is – similarly to the scaling-up effect in efficiency – a "scaling up [of] the biases and inequalities" (E8).

Limited diversity/range in human morality and perspectives. 33% of the experts stressed that formalizing ethical principles within algorithms/technology could lead to a decreased diversity in human morality and perspectives. Generally, the idea that computational ethics changes morality was considered a (rather undesirable) potential outcome (E2). For one, pressures on human morality would be created, which influences "how fast or slow moral concepts change" (see also the 'Inflexibility and lacking timeliness of technology' argument in 0) or "how much diversity there is in human moral beliefs" (E7). For example, one expert added that a focus on the values of Western, educated, industrialized, rich, and democratic (WEIRD) societies might emerge since most research studies rely on data from this sample group (E8). Thus, one challenge of computational ethics is "to pay attention to decolonizing this discourse" (E3) and the representativeness and incorporation of data and values from diverse groups (E10). Another expert pictured what would happen if values were translated/embedded into software and such technologies would then be adopted in other regions:

"Once you do it and, your software will go over the world through the internet and then end up in contexts where it's not adequate. Let's say [...] it will be on my computer [...]. My nationalism or my commitment to [my home country] is all of a sudden having an American twist. That's not what I would like, and [...] it definitely is bad if I don't have means to control it, switch it off so that I just get squeezed into my systems, the way in which values of another group are materialized." (E12)

Another expert shared the opinion that computational ethics may enforce a limited variation of acknowledged perspectives and referred to the issue of stereotyping: As particular values or experiences “are left out of the equation”, technologies can “have a very narrow focus on what is normal” (for example, “what it means to be a woman or to be typically feminine”) (E3). Apart from these direct potential effects of formalizing ethical principles into algorithms/technologies, one expert stated that the process of undertaking computational ethics itself limits the societal involvement and value considerations to those skilled in informatics and computing. At the same time, “lay people might have the feeling that they do not know how or that they don't have the capabilities to take part in the discourse” (E4).

From an *epistemic perspective*, experts indicated that computational ethics is associated with some destructive consequences for human knowledge and that there is a lack of information preventing its successful implementation:

No ground truth about what ethical principles to code. 67% of the experts communicated the existing disagreement about which ethical principles exactly to formalize within algorithms/technology. One expert named this as the key issue when it comes to computational ethics:

“The biggest challenge with codifying ethics in machines is that there is no real ground truth. So, there is no ultimate rationalization for certain ethical principles. Who really knows which principles [...] are the right principles, which [...] are the right training stimuli?” (E5)

Similarly, another expert expressed ambivalence about considering certain factors (e.g., the family status of a patient) in automated allocations of scarce medical resources by saying, “these factors should not be part of the [decision], or maybe they should, I don't know” (E8). Also, across countries, it may differ what factors are legitimate to be considered. For example, one expert referred to the importance of a patient's age in the US when making decisions about organ donations, which may be “controversial in some parts of Germany” (E6). Therefore, a few experts agreed that it is illusory that we will ever come up with “a universal ethics that we could all agree upon” (E3). More specifically, one expert said it is unrealistic that we will “have some harmonious one-world situation where everybody becomes a rule utilitarian” (E9). This creates moral uncertainty about which normative theories to draw on when engaging in computational ethics (E10). As a potential solution, one expert highlighted “the basic principle of [maximizing every individual's] choice” as a universal theory, which, however, can also lead to conflicts and contradictions (E11).

Misunderstanding/multiplicity in terminology and language dealing with computational ethics. 42% of the experts stressed that the field of computational ethics

lacks concrete language or terms that are universally understood or adopted. One expert highlighted the fuzziness and imprecision of utilized terms by saying that in the past, a distinction between 'computational ethics' and 'machine ethics' was missing and that the former is much more explanatory while the latter is rather general and "complicated" (E11). Even agreement on the term 'computational ethics' is not self-evident as "there are many terms that compete" to describe the activity "of putting ethical capability [...] inside a machine" (E9). Especially across disciplines and "between the normative and the technology communities", there are "huge gaps" in how they speak (E10). Another expert talked about the challenge of engineers talking to philosophers as they prefer diverging clarity of terms or use "different words, different jargon", even if they talk about the same issues (E3).

Limited control/verification if intended values are realized. 33% of the experts addressed the limited ability to transparently control and monitor if the ethical principles that were embedded are indeed realized as intended. One expert stated that with computational ethics, "you're ultimately leaving it up to the algorithm" since, after encoding, the algorithm will "take off from there" (E3). Therefore, it is very difficult to know and verify how the software will eventually work (E9) and whether embedded values correspond to "real-life values" (E2). As one reason for this mismatch, an expert mentioned the fact that (embedded) values can be redesigned not only by designers but also by the technology's users in that they, for example, "use the system in a different way than was intended" (E2).

Uncertainty about what constitutes (good) moral judgment/decision-making. According to only 25% of the interviewed experts, we currently have a limited understanding of what constitutes 'good' moral judgment. For example, one expert said, "we have quite naïve conceptions of morality" in that there is no "accepted theory or agreement on [...] how we make good moral judgments" or when ethical decision-making is "improved" (E2). This, in turn, leads to ignorance concerning what exactly to implement when engaging in computational ethics. Similarly, another expert raised concerns about human's ability to determine ethical principles to be encoded:

"I am not optimistic about purely top-down strategies for creating computers that reason about morality. I think our understanding of ethics, ethical principles, and our own moral reasoning and cognition is too rudimentary." (E7)

Due to our limited knowledge of ethics, one expert stated that it would be hazardous to implement something within machines "that we don't understand yet" (E11). Thus, one expert proposed, "we have to improve ourselves first" before expecting machines to create a better society (E1).

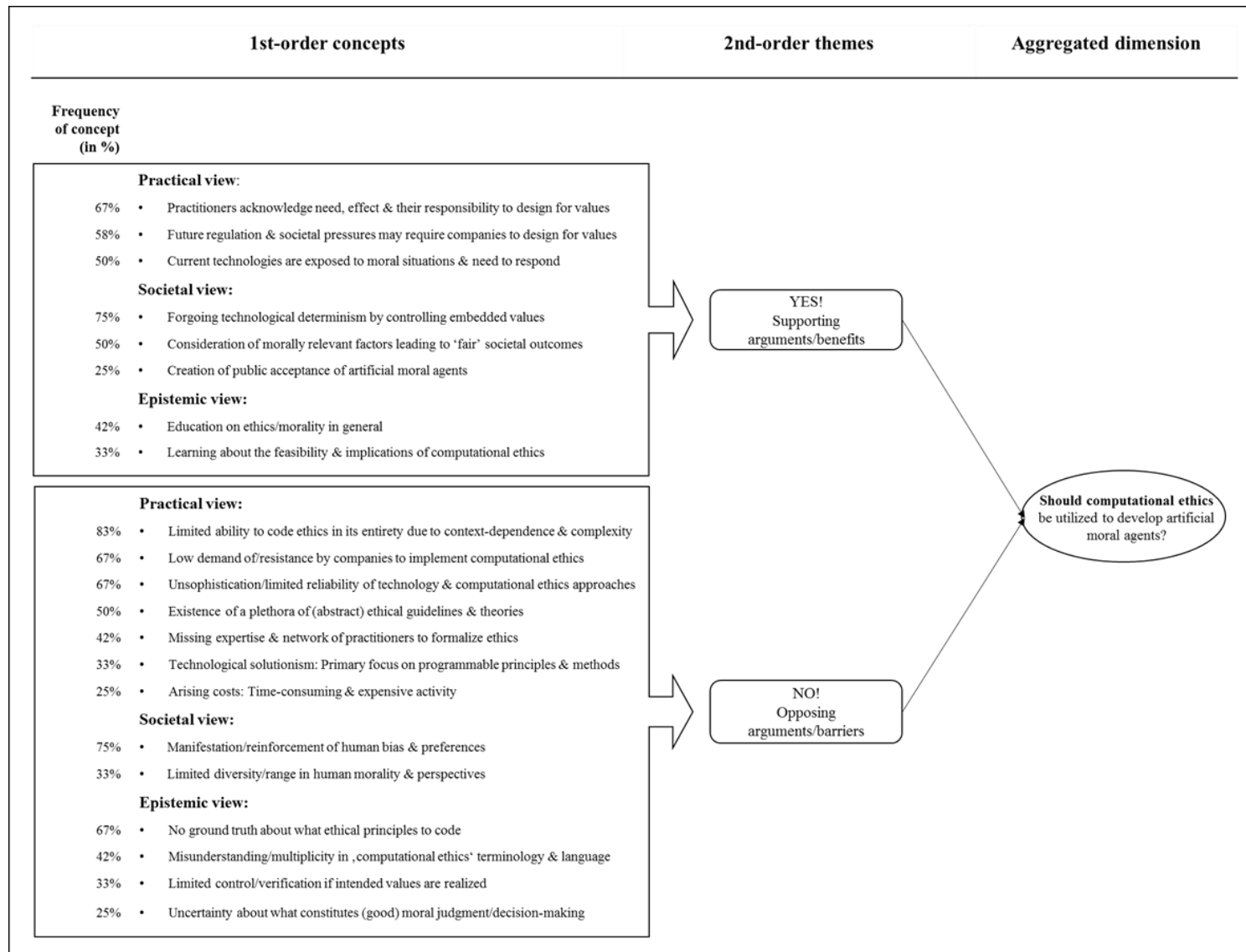


Figure 7: Experts' indicated arguments for and against utilizing computational ethics to develop artificial moral agents (ordered by perspective and frequency)

2.4.3 Recommendations for implementing computational ethics

While the former sections illustrate that experts mentioned many arguments for and against AMAs and the use of computational ethics, this section focuses on how to do so, thereby providing answers to our third research question. After all, focusing on the 'how' instead of the 'if' is much more fruitful anyway, as this expert highlighted:

„So, the consequences are all affected a lot by whether we do it properly or improperly, right? [...] The issue is not whether we can do it or should do it. But HOW it should be used once we develop computational ethics.” (E6)

Therefore, this section descriptively summarizes recommendations for implementing computational ethics and developing corresponding technology that experts indicated during the interviews (see Figure 8). These suggestions can be clustered into recommendations for the company's design and development, for industry's governance, and for scientific investigations, which will be summarized in the following.

Recommendations for company's design and development

In the interviews, experts stated suggestions for companies that (aim to) produce AMAs and attend to computational ethics. Specifically, they talked about methods to technically implement ethical principles within algorithms, what may constitute adequate underlying references, potential restrictions and limitations that should be considered, and the general workflow process of implementing computational ethics.

Technical implementation. 92% of the interviewed experts stated three different approaches to technically deriving and formalizing ethical principles within algorithms/technology, namely, it can be conducted in a top-down, bottom-up, or hybrid fashion. *Top-down approaches* constitute a method in which certain ethical principles are preprogrammed into the technology's guidance system by, for example, the developer (E2, E10) so that the technology displays “commitment to particular ethical principles, theories, codes” (E9) or to certain morally relevant factors (E6). Through *bottom-up approaches* (E4, E10), ethical principles are not directly pre-determined or programmed. Instead, technologies determine ethical principles/behavioral patterns for their guidance system themselves. In the context of computational ethics, one expert said bottom-up approaches “mean that you so to say ‘crowdsourced ethics’” (E5). For example, they do so through “deep learning approaches [by seeing] how people make judgments and try[ing] to find patterns in it” or eventually mimicking the moral reasoning and decisions of humans (E2, E6, E7). One expert highlighted the predominance of such methods by saying, “right now, deep learning is making the biggest success in many applications” (E8). One reason for this success may

be the “massive amounts of statistical data” available to train technology in such ways (E9). One expert mentioned another bottom-up technique that trains language models based on texts about different moral stories and normative decision-making (E5). Lastly, *hybrid approaches* combine top-down and bottom-up techniques, which two experts perceived as the best method (E5, E9). One expert elaborated on why both methods are needed:

“We need the bottom-up approaches in order to have a sufficient amount of training stimuli. [...] Otherwise, we will not achieve flexible systems that we can use in open-domain situations. But we combine this bottom-up approach with a top-down approach where people, who are, so to say, ethical experts, or also experts in moral psychology, filter which stimuli are allowed to become training stimuli.” (E5)

Overall, one expert said that for now, different technical methodologies should be flexibly adopted so that “every technique that we have today” can be tested to ultimately identify the most successful technique (E8).

Underlying reference points. Experts indicated who should/could be involved in deciding which ethical principles to formalize and which ethical principles exactly eventually should/could be formalized. Concerning the *stakeholders involved* in this process, 92% of the experts pointed towards the need for inclusive participation and named (ethical) experts as well as users and the broader society. As relevant people to include, interviewees highlighted groups of ethical experts (E8), such as experts in moral psychology (E5). These expert groups would then “do a first step of filtering out the things that are obviously wrong” and develop “a set of solutions that are ethically defensible” before opening up the debate for feedback, tests, or adjustments to the public (E8). Another interviewee introduced the meta criterion that “all the stakeholders, all those who are involved [...], everyone who is affected should be involved”, such as the clients of a technology (E11, E12). In this regard, one expert said that they are conducting “surveys of the general public to figure out what most people think is morally relevant” so that technologies do not “override human values” (E6). To summarize, experts emphasized “it is a joint decision” (E12), in that when “deciding what actually should happen, it should be an approach that combines both experts and the public” (E8). Concerning the *ethical principles to be implemented*, 75% of the interviewed experts referred to different normative theories/principles, legal standards, company values, or contextual factors. In terms of normative theories and principles, integrating a mix of different normative theories, such as utilitarianism, was proposed (E9, E10), amongst others, due to prevailing moral uncertainty (e.g., the ‘No ground truth about what ethical principles to code’ argument). Similarly, descriptive ethics could be an important source of information for deciding what ethical principles to implement (E11). When adopting human values within

algorithms/technologies, these should be “the real values that [individuals] really hold at a deep and profound and fundamental level” and “if they were informed, rational and impartial” (E6). Another expert referred to normative theories by stressing that human values must be complemented with and correspond to a normative foundation (E3). Furthermore, experts stated the importance of applied ethics (E4), for example, “a bioethics which emphasizes beneficence and autonomy” (E11), justice, explainability, or “no harm principles” (E2). In addition, some context-specific morally relevant factors were mentioned as crucial reference points, for example: in the case of automated allocations of scarce medical resources, these could entail the urgency/health condition and age of the patient or the success rate of an imminent operation, while factors related to ethnicity should be neglected (E6). The expert added more “controversial” personal characteristics that could be considered, such as the patient’s family status, criminal past, or own personal responsibility for their health status (e.g., due to the consumption of illegal drugs) (E6). Legal standards and regulations were also indicated as potential underlying reference points (E9, E10), “so you could probably just develop computational ethics [...] all the way up to the boundary of law” (E11). As an example, experts mentioned the acknowledgment of human rights (E4) or programming restrictions and values derived from the Geneva Conventions (e.g., necessity or proportionality) into drones (E2). Lastly, company and industry-specific codes of conduct (e.g., transparency, accountability, no paternalism) should also inform the programming of an algorithm/technology, according to two interviewed experts (E9, E12).

Potential restrictions/requirements. Experts pointed towards possible restrictions and requirements for formalizing ethical principles within algorithms/technology, such as limiting the technology's autonomy, responsibility, and operation space, as well as suggesting application-specific development of ethical frameworks. In terms of *limiting the technology's autonomy and responsibility*, 92% of the experts highlighted that pertinent technology should attend to the commands of their human users or developers so that “it does what we want it to do” (E6). In practice, this means that technologies are only automated to a certain degree (E4) without “complete replacement of humans” (E2) in the sense that, for example, humans are allowed to have the final say or to override the proposed decisions (E6, E12), by “push[ing] a button” (E2). This should be possible, especially when the user notices the technology is not functioning properly (E6) or when the technology operates in completely new situations (E8). After all, one expert stated there would always be situations in which decisions must be deferred to the human overseeing the technology’s operation (E8). On the other hand, it is also perceived critically to what extent humans are always “a very good backup” (E10) since they will “introduce a level of randomness” by their personal

decisions and actions (E11). Nevertheless, the ultimate responsibility “when something goes wrong” should remain with the humans behind the technology’s development and not pass on to the technology itself (E11) since it is not a “fully moral agent” but rather an assistant to humans (E5). Another expert talked about the link between computational ethics and the attribution of responsibility in the following way:

“[I]f there are any conditions in the near term where it would be acceptable to have a CE [computational ethics] system make a decision or generate a judgment, the cases involved will be very simple. [...] [Amongst others,] they will also not be cases where the entity that makes the decision or forms the belief is supposed to be capable of discharging obligations or bearing responsibility for their actions or beliefs.” (E7)

In addition to restricting the technology’s autonomy and responsibility, 58% of the experts pointed towards *limiting the technology’s operation space* as “a very good rule of thumb” (E10, E12). This space would then confine for which decisions, situations, or contexts ethically informed technologies can/should be utilized. Through “constrain[ing] the space where in the system can operate” could result in the system not engaging in some unethical decision or activity (E2). To do so, one expert proposed to “look at the [specific] actions before automating” and think about whether “there [are] some things that we do not want to render for automated decision-making” (E4). For example, these could entail decisions in complex or high-stake contexts (E7), such as the automated decision-making of vehicles (E11) or in “politics or health, [...] where significant physical harms can occur” (E5). Another expert proposed that computational ethics would only be accepted for decisions that “inspire little human disagreement, [...] not involve conflicting interests” (E7). Additionally, it was emphasized that restrictions should be set for automating decisions when technical robustness is not guaranteed or “where obvious technical shortcomings are detectable” (E5). Since one expert was very skeptical about the reliability of computational ethics and pertinent technology, it was suggested not to implement corresponding efforts “into the real world” or beyond experimental settings (E2). The last indicated recommendation concerning restrictions related to the *limited transferal of ethical decision-making models*, meaning that computational ethics and corresponding algorithms should always be developed specifically for a particular context and technological application (E3, E4). One expert provided an underlying reason for this:

“In the literature, it's being argued that you can only give content to a value within a certain context, and then translate it into requirements with some realized value in that specific context.” (E12)

Another expert agreed that you could not adopt the same ethical principles or decision-making procedure for different technologies by referring to the following example:

"We got to make sure that the situation that it was designed for is...you stick with that. You don't all of a sudden take something that was assigned for one thing and apply it to something else. So, for example, it'd be very simple to say: Well, we did it for kidneys. Now, let's do it for livers. [...] Let's do it for...No, No, no!" (E6)

Instead, what ethical principles are ultimately formalized within the algorithm/technology has to be contextual and depends on the use case, "whether it's autonomous cars or ethical robots, or whatever" (E11).

Workflow process. Apart from the previously stated recommendations for a company's design and development of pertinent technology, experts also made remarks about how to structure this process overall by emphasizing the importance of redress mechanisms, iterative and continuous development, comprehensive ethical reflections as well as transparent and open communication. Firstly, 83% of the experts highlighted to found *redress mechanisms* that check and adjust the technology's compliance to underlying formalized ethical principles. Especially due to the infancy of efforts in the field of computational ethics (E8), "it's important that we, as humans, keep reevaluating the sort of the ethical soundness and accurateness of that algorithm that we designed" (E3) and "take the outputs with a grain of salt" (E5). An example of such a "functional check" would be whether the embedded values are actually realized during human-technology interactions or whether a "certain input leads to the right output" (E12). Similarly, one expert said that this could be validated formally or in a statistical manner, such as:

"I gave it a 1,000 ethically charged cases of a certain case, and in 972, it did a good job. So I'm gonna say that's good enough." (E9)

Another expert pointed towards the importance of a "council of ethicists" who assess and verify if a system's ultimate decisions are ethical (E11). If discrepancies emerge or the technology does not show compliant behavior, companies should have mechanisms to fix these issues (E7). Secondly, 67% of the experts recommended implementing and rolling out computational ethics in an iterative, *step-by-step manner* instead of imposing it "too quickly or immediately" (E6). Therefore, this expert emphasized not to program all at once but to focus on a "pilot study to start with and then maybe expand from [there]" (E6). In such trials, it could then be tested to what extent ethical principles can be formalized and what decisions would result in different contexts (E6) or when using other inputs (E8). Therefore, the process of implementing computational ethics is similar to software engineering in general,

which follows an “agile” and “long process” involving beta tests to identify if anything is missing or needs adjustment (E8). Another expert expressed a similar thought by saying:

“It's going to be an iterative process...as anything in engineering, you need to take a sample first, and you know, test it and then scale it up.[...] You know, a 30% sample, seeing how it goes, making the changes, fine-tuning.” (E10)

For example, such tests could be conducted “over and over again through simulations” (E11). More generally, an iterative investigation could be conducted by testing computational ethics and pertinent technology first in decision contexts that are not emergencies or time-sensitive (E6). Once it is established that such systems can handle basic decisions (E9), computational ethics could be introduced “in more risky areas progressively”, such as autonomous driving (E11). Thirdly, 42% of the experts noted the need for developers/companies to engage in far-reaching, *‘bigger picture’ ethical reflection* when implementing computational ethics to recognize, for example, the resulting societal implications or existing limits. One expert said in this regard:

“Computational ethics is not a silver bullet, as it sometimes is argued as or presented as. And it should always be seen as secondary towards the larger framework of ethics, AI ethics, technology ethics and designing desirable technologies.” (E4)

Therefore, the expert elaborated further, “it's always good to take also a step back and ask, what is the bigger picture we are doing this in”, which, for example, includes “taking into consideration: what do we mean with the formalization of principles” (E4). Another expert shared this opinion and suggested critically engaging in a simultaneous “true ethical thought about what we are doing, [...] why we are doing that, and how are we going to assess what we have done” when implementing ethics into software (E1). Further questions for such reflections were pointed out to be able to anticipate the broader consequences of formalizing ethical principles within technology:

“Is this really the outcome that we want? Or are we indeed actually taking all important things into consideration? [...] What is the aim of the certain technology that you're talking about? Who are the people that we will be using the technology? Who benefits from it? Who doesn't? For whom is it intended for? [...] What technology users are neglected? [...] What cultural assumptions are sort of hidden into our scope?” (E3)

One expert added the importance of building up a responsible company culture by saying that such reflection processes and “ethical sensitivity [...] can only be present or can only exist if the culture promotes that” (E10). This is in line with the indications of 25% of the experts who argued ethical principles need to be considered *from the start and throughout*

the entire development process of a technology (not just when designing its algorithm). Lastly, 33% of the experts stressed *open and transparent communication* about the technology's underlying logic (i.e., ethical principles) and its limits to consumers (E1). One expert drew parallels to other products to elaborate on this recommendation:

“Just like a package of cigarettes, it should be stated: This software contains ethics. If you use this, this and this will happen, such that users have a choice of switching it on or off or using it, and it should be said what it contains in ways which are accessible for users. That's what codes of conduct for design say: it should inform your users what I bought [...] what the effects are, even if they are negative.” (E12)

Recommendations for industry's governance

Experts indicated suggestions for governing computational ethics and the industry of pertinent technologies by naming company-external enablers and complementary activities that advance its implementation in practice. These included product diversity, interdisciplinary collaboration, and education, as well as the provision of incentives.

Diversity in market supply. 17% of the experts highlighted the need to maintain diversity in the technological products offered in a market to assure consumers' freedom of choice. One expert emphasized this by saying:

“If you go [...] to a supermarket there, you [...] have a freedom of choice. Namely, you have lots of beers, lots of brands, and lots of offer. [...] If you translate that to technology: [...] Some have a car like this, others have a car like that. [...] So, if someone offers computational ethics but everyone has to follow it because we don't have another choice, [...] then we have a problem that we are going straight back to the seventh century, sixteenth century that there was a State Church and the ones who would have a different background doomed.” (E12)

Therefore, to prevent the establishment of a monopoly (especially in terms of a particular ethical decision logic), there needs to be “serious opt-out possibilities” and alternative products that are available in a market (E12).

Interdisciplinary collaboration and education. Experts expressed the importance of interdisciplinarity within cooperation and educational programs to facilitate computational ethics and establish a joint language in this field. To elaborate, the first recommendation entailed engaging in *interdisciplinary, multi-stakeholder collaboration*, as such a working style is “quite self-provoked” due to the nature of computational ethics (E4), which requires knowledge “in computers as well as in ethics” (E6). Therefore, it was suggested that team members developing pertinent technology are not trained in the same discipline but are, for

example, philosophers, data scientists, computer scientists, policy specialists, and experts in the particular use case/domain (E4, E6). Having “practitioners from that field that have this built-in knowledge from the domain” is vital in order not to lose any implicit knowledge that these practitioners pass on by “intuition” (E10). Also, among research groups, the need for interdisciplinary team compositions was stressed (E7, E9). For example, one expert talked about their own interdisciplinary undertakings in the future to be able to apply theoretical ideas:

“Currently, I am doing only theoretical work, but I hope in the future to collaborate with computer scientists on projects that attempt to model or simulate human moral reasoning.” (E7)

Another expert said it is helpful for researchers to work together with practitioners “to step out of the bubble and see what's the real situation at the moment” (E4). In fact, engineers, designers, and philosophers from the research community struggle with similar issues and questions, so the potential for synergies exists, according to one expert (E3). As a side effect, more collaboration among disciplines could also help individuals better understand their different perspectives and jargon (E3, E11). *Determining a (joint) explicit language and vocabulary* for discussing computational ethics was indicated as another critical recommendation. For example, one expert emphasized clarifying and uniformly using the term ‘autonomous’ and refraining from “anthropomorphisms when talking about machines” (E1). Another expert expressed uncertainty about “talk[ing] about advanced computer systems that generate utterances containing moral terms” and mentioned that describing them as ‘ethical’ is inadequate (E7). A prominent recommendation entailed the *‘ethical’ education of programmers* who develop pertinent technology. Thus, in addition to ensuring “the AI itself is artificially making ethical decisions”, the focus should be placed on the “designers, the people designing these technologies” (E3, E5) so that “ethically important points” can be recognized (E4). This can be done by providing guides to programmers on “how to better engineer AI” (E3, E9), meaning what principles or values to keep in mind during the design and development phase (E1, E2). However, “having a check box in the design phase is not enough”, according to one expert (E10). Another expert shared this opinion when saying that “just writing a few principles on a piece of paper and hanging it on the wall” is ineffective (E5). Instead, experts emphasized that there has to be a shift in the educational system toward offering courses in the field of computational ethics:

“You have to look at what you're teaching. You have to change what you're teaching and [...] from a government standpoint, you have to insist on some things being taught in order for this to be addressed.” (E9)

Another expert provided a more explicit example of what “robust ethics training” at universities comprises:

“It cannot be just an ethics course that is taught at the engineering level. It needs to be an engineering course that really incorporates ethics.” (E10)

Through such new courses, engineers and developers would learn about and come to understand the ethical implications of technology (E10). Such education has to reach beyond the university level by continuing to train specific behavioral dispositions and virtues of employers in companies directly (E5, E10). Overall, implementing ethics into algorithms (i.e., computational ethics) and educating responsible parties to be able to do so were perceived as relevant “complements” to each other (E3).

Incentives and controls. Experts recommended creating incentives and (regulatory) measures for practitioners in the field so they are prone to attend to computational ethics. First, 33% of the experts emphasized the need to *establish and enact standards and regulatory measures* that provide a universal framework and more guidance for industry. For example, companies may only be more receptive and accepting of computational ethics once “the ISO put out an ethical machines code that said [...] all machines have to follow these certain ethical principles” (E11). However, it is important to strike a balance between regulation and self-governance of companies to allow the emergence of creative processes during algorithm design (E3). 17% of the experts recommended *offering (financial) support for companies and researchers* to pursue the formalization of ethical principles within their algorithms/technologies. According to this expert, without a “fertile funding landscape” and any “government-sponsored programs for R&D” in the fields of computer science and AI, it will not be possible to “DO the formalization and to DO the implementation at the level of algorithms and beyond into machines” (E9). Another expert also mentioned that such funding should ideally come from “the government or non-profits so that research on the topic [...] can proceed in the absence of pressure to push under-developed and under-tested systems into the world” (E7). Lastly, one expert stated the need to *highlight technology's ethical implications* to technology companies and, thereby, create moral awareness around computational ethics. The expert said if companies were more informed about the ramifications of their products and that they pose “moral problems”, some would actively start engaging in formalizing ethical principles within algorithms/technology “even though it’s going to cost them money” (E6).

Recommendations for scientific investigations

Experts indicated suggestions for the scientific community in the field of computational ethics by stating what kind of research should be focused on and what open questions to investigate in the future.

Research focus. Concerning focal points in the inquiry of computational ethics, the interviewed experts pointed towards appropriate scientific methodologies and processes for advancing investigations. First, 83% of the experts stressed the *simultaneous engagement in theoretical and empirical studies* for encountering adequate ethical principles to be formalized within algorithms/technology. Fundamental, “rigorous” understanding and research on (embedding) ethical reasoning (E2, E9) as well as on systems that are based on computational ethics is necessary (E7). In addition, empirical research is key (E7) to, for example, see what the identified theoretical ideas mean in practice (E2). As a more specific methodology of empirical investigations, one expert mentioned “running simulations” (E11). Such empirical investigations could then generate insights into outcomes of human-technology interactions (E4, E10, E12), such as the impact of pertinent technology on human decision-making or into the technology’s level of sophistication in making moral judgments (E2). Second, – with regard to the plethora of (abstract) ethical guidelines and theories – it was emphasized that academia needs to *concretize and formalize ethical principles*. Instead of focusing on “too general” work, experts recommended that the scientific community needs to provide practical deliverables, for example, by turning relevant ethical principles into operationalized form (E10). Other experts stated that higher levels of concretization could be achieved through applying general ethical principles in specific contexts or for specific applications (E3, E8). For example, one expert added in this regard:

“‘Don’t harm’ is something, and of course, this is too abstract still. But depending on the application, you can become more specific about it.” (E2)

Lastly, 25% of the experts highlighted the importance of developing a supportive, collaborative, and open-minded (discussion) culture within the research community. After all, one expert mentioned that the establishment of ethical principles or frameworks follows a successive development in that research builds on previous work:

“[E]thical principles [...] [and] the normative theories... They are not just put together by one person. [...] It may have been written by one person, but then it gets built on by the next person, the next person, the next person.” (E11)

Furthermore, it was suggested “to decolonizing the discourse” (E3) and the representativeness and incorporation of data and values from diverse groups (E10), for

example, by extending research in the field of machine ethics beyond Western-centric ethics and school of thoughts (E3).

Open questions. Experts gave a list of open questions that need to be studied in future research, which span from questions concerning the (technical) implementation, underlying reference points, purpose and outcomes of computational ethics, important restrictions, human ethical decision-making and its relation to computational ethics, reliability and assessment or responsibility and accountability issues as well as its societal acceptance. The specific questions are itemized in the following.

(Technical) implementation

- What is the best technical approach to formalizing ethics?
- How and to what extent can we formalize (abstract) ethical principles within the code?

Experts highlighted conducting more research to find out whether and how implementing ethics into algorithms is at all feasible (E2, E3, E6) and “physically possible” (E9). In particular, it is important to establish what ethical principles or theories are most suitable (in which contexts) (E1, E3) or which moral concepts cannot be adopted (E7). Investigations should also address to what extent embedded values within technologies should align with human values and how to generate these values (E6). In this regard, one expert said that human decision-making should not simply be duplicated within technology, but more nuanced approaches need to be researched:

“So engineers and researchers have to find out what can be implemented in a machine that is not a copy of what human beings would do in such a situation. It's like a plane that is not like a flying boat.” (E1)

Underlying reference points

- What are the underlying reference points for computational ethics?

Future investigations need to address which (abstract) factors to include when formalizing ethics within algorithms and how to achieve this translation (E9, E12). For example, it is unclear “what kind of consequences” should be accounted for (e.g., “immediate consequences, the consequences of the consequences, for whom?”) (E1) or “what the formal definition of fairness” is (E9). Furthermore, research has to focus on examining “what sources can be used to gather training stimuli” (E5) and what are adequate datasets (E1, E10). Furthermore, it needs to be determined “what labor allow[s] you to produce” a translation of ethical principles into an operationalized form (E9), so it needs to be established who is involved in the process.

Purpose and outcomes

- For what purpose is computational ethics needed and desired?
- What are the benefits of computational ethics and of technologies engaging in ethical decision-making?
- What are the limits of computational ethics and of technologies engaging in ethical decision-making?

Experts indicated the importance of investigating the “ultimate aim” of computational ethics and if this “is actually what we want? And why?” (E3). Another expert stressed that “the first question that we should ask ourselves” is about its purpose and expected results “before try and implement principles and ethics into machines” (E1). In this regard, it is helpful to understand “what kind of societies we would like to live in and how could these technologies help us achieve those societies” (E4). Therefore, research needs to address “what computers can do on their own, how far they can get with moral judgment”, how well they work in collaboration with humans (E2), or “what is lost in the process” (E4).

Restrictions

- What are areas where artificial moral agents should not be offered at all?
- What are situations in which humans should not be able to override the system’s decision?

Experts pointed towards open questions that address “where we should usher in technology and where we should exempt it” (E1, E5) by investigating if “there are some things that we do not want to render for automated decision-making” (E4). Similar to this general consideration, it is questionable when humans need or “ought to be able to override” a technology’s decision (E6).

Human ethical decision-making and its relation to computational ethics

- How do we capture ‘true’ human values?
- What is the relation between the ethical decision-making of humans and technology?

42% of the experts highlighted the need to further establish our understanding of human morality and “what’s the real basis” for human ethical judgments (E1). Furthermore, it needs to be investigated how to compare “computational with human decision-making” (E1).

Reliability and assessment

- How can the proper functionality and compliance of technology be assessed post-hoc?

Furthermore, future research needs to address “how [we] are going to assess the results” and who and “what are the references to say that [an] effect is positive or negative” (E1). Open

questions also remain concerning how to test “whether values really have been embedded in a technology” (E2) and how to ensure no blind spots or prejudices are adopted (E3).

Responsibility and accountability

- How do we manage emerging responsibility gaps?

17% of the experts stressed the open question of whether companies or machines themselves should be considered responsible parties and how emerging responsibility gaps can be managed (E7, E11).

Societal acceptance

- To what extent will AMAs be accepted by society?
- To what extent can we agree on underlying ethical principles universally?

One expert stated it is currently unclear if the corresponding technology will be well received by society in the future (E5), pointing towards necessary investigations about social acceptance and adoptions. Similarly, it is questionable if “there is a universal ethics that we could all agree upon” that is formalized within algorithms/technologies (E3). Another indicated open questions related to the generalizability of particular ethical principles or decision logic across various technological applications (E6).

Essay I: Formalizing Ethical Principles Within AI Systems: Experts' Opinions on Why (Not) And How to do it

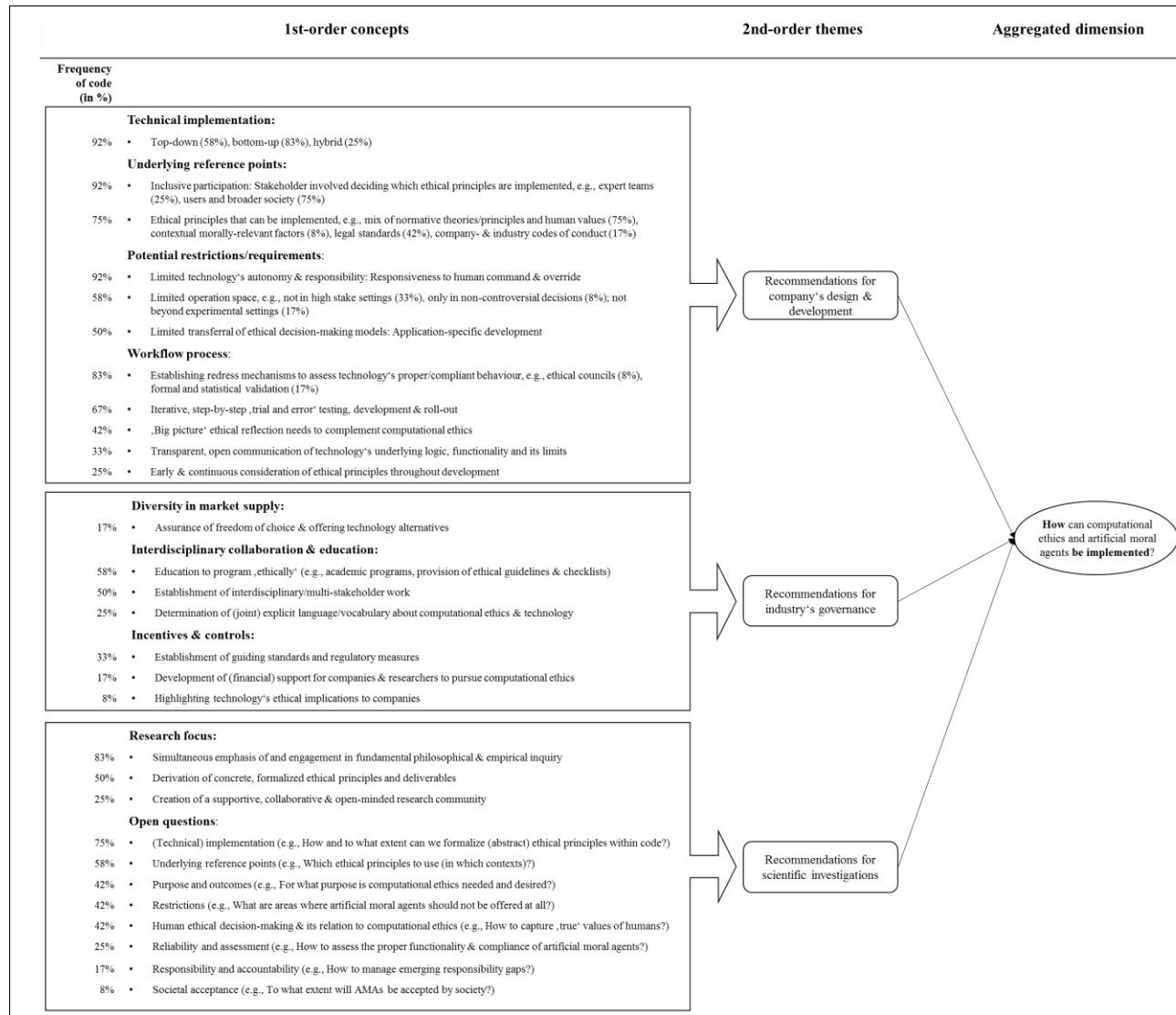


Figure 8: Experts' indicated recommendations of how to implement computational ethics and artificial moral agents (ordered by type of recommendation and frequency)

2.5 Discussion: Facilitating computational ethics and artificial moral agents

In this section, we briefly construe three key learnings and derive a model for facilitating computational ethics and AMAs (see Figure 9) that aims to integrate the previously stated findings.

Computational ethics as an integral part of ‘ethics in design’. When we asked experts in our interviews about reasons for and against computational ethics, they immediately also provided reasons for and against developing/offering AMAs for ethical decisions (see 2.4.1). Therefore, it seems the method of computational ethics cannot be evaluated without considering its resulting technologies. Instead, computational ethics is one integral part of the ethical design of corresponding technology. As illustrated in Figure 9, companies need to engage in activities that relate to the programming of the system’s internal code or algorithm (e.g., technical investigation or formalization and computing) and in those that relate to the ethical design and use of the AI system overall (e.g., scope investigation, testing, and experimentation or communication). We therefore show (how) the two approaches, ‘ethics in design’ and ‘ethics by design’ that were introduced in past literature (compare 2.2), need to be considered in combination.

A mix of company activities and external enablers. For the effective execution of computational ethics, the activities of technology companies are shaped by and can be facilitated through government initiatives as well as educational and scientific efforts. Concerning necessary *company activities*, we propose their process for developing AMAs can be summarized and structured as follows. Companies need to engage in a ‘technical investigation’ to determine what approach (e.g., top-down or bottom-up) to consult when programming the system. Furthermore, they need to carry out an ‘ethical principles investigation’, in which it is established what ethical principles are utilized as underlying/embedded reference points (e.g., normative theories or legal standards) and who (e.g., ethical experts or the broader public) decides on these. Based on these insights, the ethical decision-making logic of the AMA can be formalized and computed. In addition to this programming, companies can conduct a ‘scope investigation’ to determine whether there should be any restrictions to the operation of the algorithm and the AMA overall. Then, companies can test their functionality before the technology is put on the market. Simultaneously, companies should engage in impact assessments and redress activities to ensure the AMA operates as planned (e.g., realized values equal intended/embedded values). In addition, companies need to communicate the technology’s underlying logic, functionality, and limits to their users. *External enablers* can help companies achieve these previously indicated activities. According to our findings, such enablers are, for example, the assurance of ‘diversity in the market supply’, ‘interdisciplinary collaborations and education’, the provision of ‘(financial)

incentives' as well as further 'scientific investigations of open questions'. More explicit explanations and first recommendations of what exactly could be contemplated at every single step of this proposed model can be viewed in Section 2.4.3 and Figure 8. Therefore, this study offers a structured and more detailed framework outlining factors to be taken into account at various stages of AMA development and extends relevant considerations (e.g., underlying reference points) that were highlighted in the theoretical background section.

The importance of 'bigger picture' reflections. As implied by the numerous and (at times) dominating arguments from a societal and epistemic view (compare Figure 10 and Figure 13 in Appendix B of Essay I – Frequency analyses of the experts' arguments & recommendations), it is crucial to engage in reflections beyond mere technical and practical aspects in the process of developing AMAs. In this study, experts even explicitly indicated developing AMAs and relying on computational ethics should be complemented with comprehensive ethical reflections. We suggest the here-discovered reasons for and against AMAs (see 2.4.1) and computational ethics (see 2.4.2) can serve as a checklist for these reflections. For example, when a company engages in its impact assessment, it could draw on the positive and negative arguments or potential outcomes that were indicated in this study and assess to what extent these ultimately manifest. More specifically, it could be evaluated to what extent the consideration of morally relevant factors indeed leads to 'fair' societal outcomes or whether this generates (physical) harm to users. Similarly, the here-discovered reasons for and against AMAs (see 2.4.1) and for and against the method of computational ethics (see 2.4.2) can serve as the foundation that justifies and invokes the establishment of certain company-internal or external activities. For example, to forego that there will be limited diversity in human morality and perspectives, the industry could push for diversity in market supply. Another example would be: To counteract users' overreliance and trust in a particular technology, companies could clearly communicate the technology's functionality and limits. Overall, merging the proposed implementation strategies (see 2.4.3) with the indicated supporting and opposing arguments (see 2.4.1 and 2.4.2) can result in a more reflected design of AMAs, which may ultimately turn these AI systems into a force for good as desired.

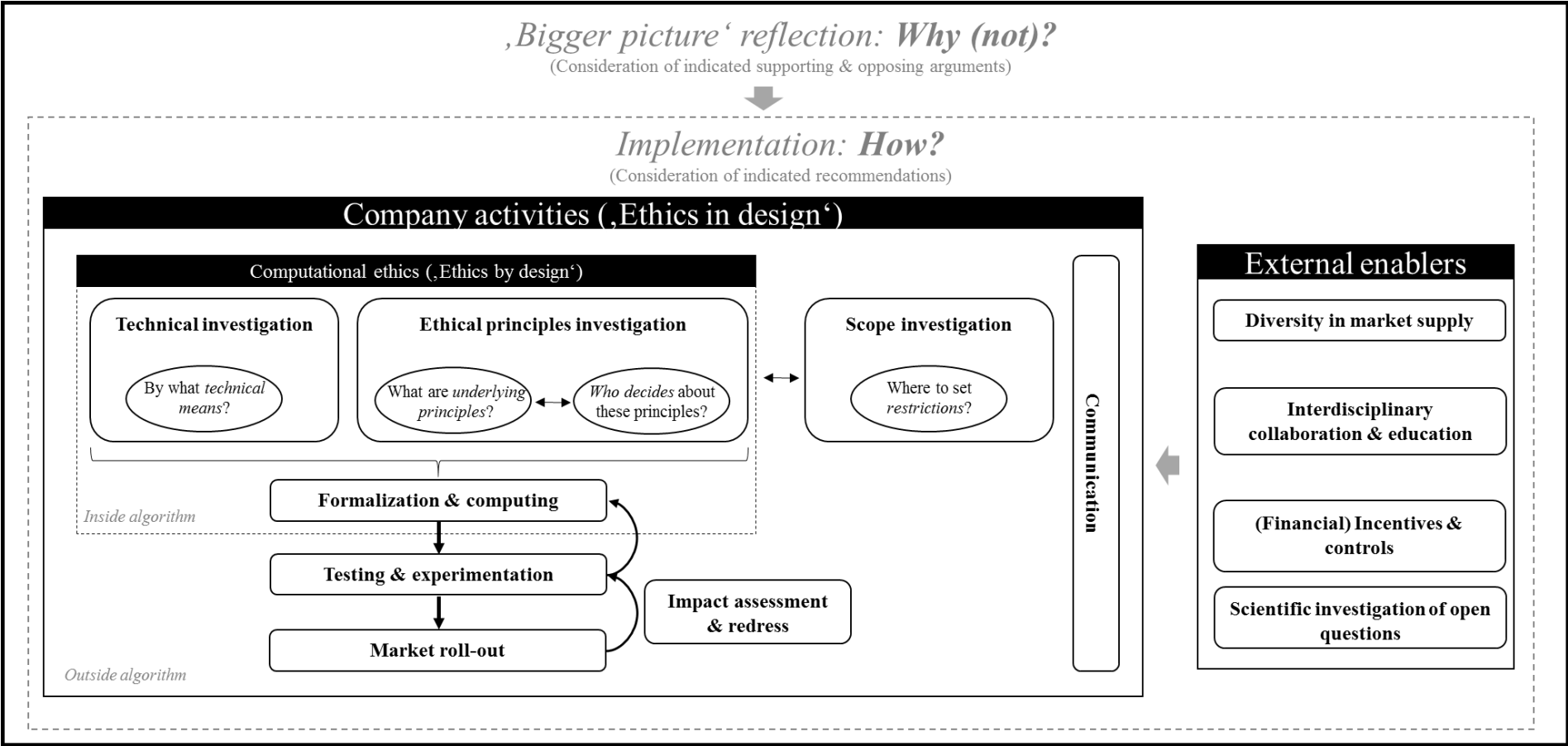


Figure 9: Proposed model for facilitating computational ethics and artificial moral agents

2.6 Conclusion

This study centers on the topic of artificial moral agents and the method of computational ethics. Based on qualitative interviews with experts in the field of philosophy, AI, and cognitive sciences, the study explores reasons for and against encoding ethical principles into algorithms/AI systems as well as for and against resulting technologies (i.e., AMAs). Findings suggest that indicated arguments can be clustered into pragmatic/practical, societal, and epistemic reasons. Particularly, with the last two types of reasons, this study moves existing debates beyond a technical perspective and provides important (ethical) considerations that need to be contemplated and governed in the context of computational ethics and AMAs. Furthermore, indicated recommendations on how to realize computational ethics and how to develop AMAs have been recapitulated. After all, experts highlighted it is more important to start thinking about the 'how' instead of devoting time to the more trivial question of whether we should do it at all. To contribute to the investigation of 'how to do it', this study proposes a model for administering computational ethics and AMAs that summarizes important company activities when developing pertinent technology as well as relevant external enabling activities by the state or educational/scientific institutions. In the future, this rather general model could be applied to a specific use case (i.e., when developing a specific technological application) and tested empirically. Despite its missing empirical validation at this point, our model and this study overall hold important implications for multiple stakeholders. It provides technology companies with a high-level model of how to structure their 'ethics in & by design' process and a list of 'bigger picture' reflections that need to be considered during this process. In addition, this study can serve policymakers to establish industry measures for governing the development of pertinent technology and points towards outstanding research questions that can be addressed by the scientific community in the future.

References

- Alt, R., Göldi, A., Österle, H., Portmann, E., & Spiekermann, S. (2021). Life Engineering: Towards a new discipline. *Business & Information Systems Engineering*, 63, 191-205. <https://doi.org/10.1007/s12599-020-00680-x>
- Awad, E., & Levine, S. (2020). *Why we should crowdsource AI ethics (and how to do so responsibly)*. Retrieved from: <https://behavioralscientist.org/why-we-should-crowdsource-ai-ethics-and-how-to-do-so-responsibly/>
- Awad, E., Levine, S., Anderson, M., Anderson, S. L., Conitzer, V., Crockett, M. J., Everett, J. A., Evgeniou, T., Gopnik, A., Jamison, J. C., Kim, T. W., Liao, S. M., Meyer, M. N., Mikhail, J., Opoku-Agyemang, K., Borg, J. S., Schroeder, J., Sinnott-Armstrong, W., Slavkovik, M., et al. (2022). Computational ethics. *Trends in Cognitive Sciences*, 26(5), 388-405. <https://doi.org/10.1016/j.tics.2022.02.009>
- Bogner, A., Littig, B., & Menz, W. (2014). *Interviews mit Experten: eine praxisorientierte Einführung*. Springer-Verlag.
- Bonnemains, V., Saurel, C., & Tessier, C. (2018). Embedded ethics: some technical and ethical challenges. *Ethics and Information Technology*, 20(1), 41-58. <https://doi.org/10.1007/s10676-018-9444-x>
- Bringsjord, S., Arkoudas, K., & Bello, P. (2006). Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems*, 21(4), 38-44. <https://doi.org/10.1109/MIS.2006.82>
- Bryda, G., & Costa, A. P. (2023). Qualitative Research in Digital Era: Innovations, Methodologies and Collaborations. *Social Sciences*, 12(10), 570. <https://doi.org/10.3390/socsci12100570>
- Coggins, T.N., & Steinert, S. (2023). The seven troubles with norm-compliant robots. *Ethics and information technology*, 25, 29. <https://doi.org/10.1007/s10676-023-09701-1>
- Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y., Kramer, M. (2017). *Moral decision making frameworks for artificial intelligence*. Available at: www.aaai.org. <https://doi.org/10.1609/aaai.v31i1.11140>
- Corbin, J., & Strauss, A. (1994). Grounded theory methodology. *Handbook of qualitative research*, 17, 273-285.
- Demir-Kaymak, Z., Turan, Z., Çit, G., & Akyaman, S. (2024). Midwifery students' opinions about episiotomy training and using virtual reality: A qualitative study. *Nurse Education Today*, 132, 106013. <https://doi.org/10.1016/j.nedt.2023.106013>
- Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Springer Nature. <https://doi.org/10.1007/978-3-030-30371-6>
- Dyoub, A., Costantini, S., & Lisi, F. A. (2020). Logic programming and machine ethics. *arXiv preprint arXiv:2009.11186*. <https://doi.org/10.48550/arXiv.2009.11186>
- Etzioni, A., & Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *The Journal of Ethics*, 21(4), 403-418. <https://doi.org/10.1007/s10892-017-9252-2>

Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International journal of qualitative methods*, 5(1), 80-92. <https://doi.org/10.1177/16094069060050010>

Floridi, L. (2023). *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*. Oxford University Press.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Lütge, C., Madelin, R., Pagallo, U., Rossi, F., et al. (2018). An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28, 689-707. <https://doi.org/10.1007/s11023-018-9482-5>

Freedman, R., Borg, J. S., Sinnott-Armstrong, W., Dickerson, J. P., & Conitzer, V. (2020). Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence*, 283, 103261. <https://doi.org/10.1016/j.artint.2020.103261>

Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: Shaping technology with moral imagination*. Mit Press.

Gehman, J., Glaser, V. L., Eisenhardt, K. M., Gioia, D., Langley, A., & Corley, K. G. (2018). Finding theory–method fit: A comparison of three qualitative approaches to theory building. *Journal of Management Inquiry*, 27(3), 284-300. <https://doi.org/10.1177/1056492617706029>

Geisslinger, M., Poszler, F., Betz, J., Lütge, C., & Lienkamp, M. (2021). Autonomous driving ethics: From trolley problem to ethics of risk. *Philosophy & Technology*, 34, 1033-1055. <https://doi.org/10.1007/s13347-021-00449-4>

Geisslinger, M., Poszler, F., & Lienkamp, M. (2023). An ethical trajectory planning algorithm for autonomous vehicles. *Nature Machine Intelligence*, 5(2), 137-144. <https://doi.org/10.1038/s42256-022-00607-z>

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2012). Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology. *Organizational research methods*, 16 (1), 15–31. <https://doi.org/10.1177/1094428112452151>

Goldstein, K. (2002). Getting in the door: Sampling and completing elite interviews. *PS: Political Science and Politics*, 35(4), 669-672. <https://doi.org/10.1017/S1049096502001130>

Govindarajulu, N. S., & Bringsjord, S. (2017). On automating the doctrine of double effect. *arXiv preprint arXiv:1703.08922*. <https://doi.org/10.48550/arXiv.1703.08922>

Guest G., Namey, E., & Chen, M. (2020). A simple method to assess and report thematic saturation in qualitative research. *PLoS One*, 15(5), e0232076. <https://doi.org/10.1371/journal.pone.0232076>

Guy, M., Blary, A., Ladner, J., & Gilliaux, M. (2021). Ethical Issues Linked to the Development of Telerehabilitation: A Qualitative Study. *International Journal of Telerehabilitation*, 13(1). <https://doi.org/10.5195/ijt.2021.6367>

Häußermann, J. J., & Lütge, C. (2022). Community-in-the-loop: Towards pluralistic value creation in AI, or—why AI needs business ethics. *AI and Ethics*, 2, 1-22. <https://doi.org/10.1007/s43681-021-00047-2>

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?. *Behavioral and brain sciences*, 33(2-3), 61-83.

IEEE (2019). *Ethically aligned design*. Retrieved from: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

Jacobs, N., & Hultgren, A. (2021). Why value sensitive design needs ethical commitments. *Ethics and information technology*, 23(1), 23-26. <https://doi.org/10.1007/s10676-018-9467-3>

Jacobs, N., & IJsselsteijn, W. (2021). Bridging the Theory-Practice Gap: Design-Experts on Capability Sensitive Design. *International Journal of Technoethics*, 12(2), 1-16. <https://doi.org/10.4018/IJT.2021070101>

Johnson, D. G., & Verdicchio, M. (2023). Ethical AI is Not about AI. *Communications of the ACM*, 66(2), 32-34. <https://doi.org/10.1145/3576932>

Martinho, A., Kroesen, M., & Chorus, C. (2021a). Computer Says I Don't Know: An Empirical Approach to Capture Moral Uncertainty in Artificial Intelligence. *Minds and Machines*, 31(2), 215-237. <https://doi.org/10.1007/s11023-021-09556-9>

Martinho, A., Poulsen, A., Kroesen, M., & Chorus, C. (2021b). Perspectives about artificial moral agents. *AI and Ethics*, 1(4), 477-490. <https://doi.org/10.1007/s43681-021-00055-2>

Mayring, P. (2014). *Qualitative content analysis: theoretical foundation, basic procedures and software solution*. Klagenfurt. <https://nbn-resolving.org/urn:nbn:de:0168-ssolar-395173>

Meier, L. J., Hein, A., Diepold, K., & Buyx, A. (2022). Algorithms for Ethical Decision-Making in the Clinic: A Proof of Concept. *The American Journal of Bioethics*, 22(7), 1-17. <https://doi.org/10.1080/15265161.2022.2040647>

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 1-21. <https://doi.org/10.1177/2053951716679679>

Moor, J. H. (1995). Is ethics computable?. *Metaphilosophy*, 26(1/2), 1-21.

Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., & Floridi, L. (2021). Operationalising AI ethics: barriers, enablers and next steps. *AI & SOCIETY*, 38, 1-13. <https://doi.org/10.1007/s00146-021-01308-8>

Moser, A., & Korstjens, I. (2018). Series: Practical guidance to qualitative research. Part 3: Sampling, data collection and analysis. *European journal of general practice*, 24(1), 9-18. <https://doi.org/10.1080/13814788.2017.1375091>

Nallur, V., Dennis, L., Bringsjord, S., & Govindarajulu, N. S. (2023). A Partially Synthesized Position on the Automation of Machine Ethics. *Digital Society*, 2(2), 14. <https://doi.org/10.1007/s44206-023-00040-8>

Nath, R., & Sahu, V. (2020). The problem of machine ethics in artificial intelligence. *AI & society*, 35(1), 103-111. <https://doi.org/10.1007/s00146-017-0768-6>

Nyholm, S. (2023). *This is technology ethics: An introduction*. John Wiley & Sons.

Phillips-Wren, G. (2012). AI tools in decision making support systems: a review. *International Journal on Artificial Intelligence Tools*, 21(02), 1240005. <https://doi.org/10.1142/S0218213012400052>

Pole, K. (2007). Mixed method designs: A review of strategies for blending quantitative and qualitative methodologies. *Mid-Western Educational Researcher*, 20(4), 35-38.

Portmann, E., & D'Onofrio, S. (2022). Computational ethics. *HMD Praxis der Wirtschaftsinformatik*, 59(2), 447-467. <https://doi.org/10.1365/s40702-022-00855-y>

Reinecke, J., Arnold, D. G., & Palazzo, G. (2016). Qualitative methods in business ethics, corporate responsibility, and sustainability research. *Business Ethics Quarterly*, 26(4), xiii-xxii. <https://doi.org/10.1017/beq.2016.67>

Salo-Pöntinen, H. (2021). AI Ethics-Critical Reflections on Embedding Ethical Frameworks in AI Technology. In *International Conference on Human-Computer Interaction* (pp.311-329). Springer. https://doi.org/10.1007/978-3-030-77431-8_20

Segun, S. T. (2021). From machine ethics to computational ethics. *AI & SOCIETY*, 36(1), 263-276. <https://doi.org/10.1007/s00146-020-01010-1>

Silverman, D. (2015). *Interpreting qualitative data*. Sage.

Sinnott-Armstrong, W., & Skorbun, J. A. (2021). How AI can AID bioethics. *Journal of Practical Ethics*, 9(1). <https://doi.org/10.3998/jpe.1175>

Spiekermann, S., & Winkler, T. (2022). Value-Based Engineering With IEEE 7000. *IEEE Technology and Society Magazine*, 41(3), 71-80. <https://doi.org/10.1109/MTS.2022.3197116>

Taddeo, M. (2019). Three ethical challenges of applications of artificial intelligence in cybersecurity. *Minds and Machines*, 29, 187-191. <https://doi.org/10.1007/s11023-019-09504-8>

Tajalli, P. (2021). AI ethics and the banality of evil. *Ethics and Information Technology*, 23(3), 447-454. <https://doi.org/10.1007/s10676-021-09587-x>

Todorovski, L. (2023). Introduction to Computational Ethics. In A. Završnik, & K. Simončič (Eds.), *Artificial Intelligence, Social Harms and Human Rights* (pp.161-179). Springer International Publishing. https://doi.org/10.1007/978-3-031-19149-7_7

Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., & Bernstein, A. (2020). Implementations in machine ethics: A survey. *ACM Computing Surveys (CSUR)*, 53(6), 1-38. <https://doi.org/10.1145/3419633>

Van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. *Minds and Machines*, 30(3), 385-409. <https://doi.org/10.1007/s11023-020-09537-4>

Whittemore, R., Chase, S. K., & Mandle, C. L. (2001). Validity in qualitative research. *Qualitative health research*, 11(4), 522-537. <https://doi.org/10.1177/104973201129119299>

Woodgate, J., & Ajmeri, N. (2022). Principles for Macro Ethics of Sociotechnical Systems: Taxonomy and Future Directions. *arXiv preprint arXiv:2208.12616*. <https://doi.org/10.48550/arXiv.2208.12616>

Appendix A of Essay I – Interview guide



Interview guide

Formalizing ethical principles within algorithms/technology

Interviewer: _____

Date: _____

Start: _____

End: _____

ID of interviewed expert: _____

Short instructions:

1. This interview is conducted as part of the study “Formalizing ethical principles within algorithms/technology” by the Chair of Business Ethics and the Institute for Ethics in Artificial Intelligence (Prof. Dr. Christoph Lütge) at the Technical University of Munich in cooperation with the Human-IST Institute (Prof. Dr. Edy Portmann) at the University of Fribourg.
2. This interview will center on computational ethics. In particular, we will ask about your personal experience with it, (ethical) reasons for and against it as well as its implementation.
3. For documentation and analysis purposes, we would like to record the following interview.
4. If you agree to participate in this study, please sign the declaration of consent.



Aims, Significance and Scope of the Study

“[A]utonomous machines, i.e. machines equipped with automated decisions functions, are intended to be put in contexts where computed decisions have to be guided by ethical considerations“ (Bonnemains et al., 2018; pp. 3-4). For example, algorithmic advisory systems are being developed that aim to assist medical professionals in critical treatment decisions (Meier et al., 2022) or an autonomous vehicle has to decide which traffic participant to target when an accident cannot be avoided (Bonnemains et al., 2018). Therefore, questions that arise are: How can we program such AI systems to be ethical? How best might we ground their ethical decisions on normative ethical principles? (Segun, 2021).

To integrate ethical decision-making processes within such autonomous machines, several approaches have been proposed (Dyoub et al., 2020). Past scholars have mentioned that these approaches need to be informed by human values and grounded in ethical theory (Jacobs & Hultgren, 2021). In order to be truly practical, these moral theories “must provide more than vague, general criteria; [t]hey must also provide an operationalizable, and presumably quantitative theory” (Conitzer et al., 2017; p.4831). In this regard, scholars have suggested computational ethics as an important frontier to move away from purely theoretical discussion towards actually building ethical AI. Computational ethics aims to translate ethical principles into computer codes to ensure that the corresponding technology is both compliant with an ethical system and functionally dynamic to operate in the real world (Segun, 2021).

Many outstanding questions concerning (long-term) consequences and the concrete implementation of computational ethics still prevail (Awad et al., 2022). Next to such technical matters, it is questionable whether ethics is and should be amendable to programming at all (Moor, 1995). Therefore, “the greater challenge in the field of AI ethics is a philosophical rather than an engineering or technological one” (Tajalli, 2021; p.448). To shed light on and provide answers to this matter, a series of qualitative expert interviews will be conducted.

Key concepts

Computational ethics: “seek[s] ways to translate abstract [ethical] principles into computer codes [...] [and to] develop ethically grounded algorithms and codes to be able to execute them during simulations and tests” (Segun, 2021; pp.20-22)

In this interview, we will use ‘computational ethics’ and “Formalizing ethical principles within algorithms/technology” as synonyms.

What is translated/implemented?

- Corresponding scholarly work aims to formalize, for example, descriptive ethics and normative ethics (Awad et al., 2022)
- Descriptive ethics: “determining what people think is morally right or wrong, and [...] formalizing these views [...] in computational terms” (Awad et al., 2022; p.396)
- Normative ethics, for example:
 - *Utilitarianism:* “An act is ethical if it maximizes the total expected utility across all who are effected” (Woodgate & Ajmeri, 2022; p.22) → operationalization by, e.g., programming an autonomous vehicle to calculate consequences for all possible trajectories and to select the one trajectory that entails the least overall estimated harm for all traffic participants (Geisslinger et al., 2021)



How is it translated/implemented?

For example, through:

- **Top-down implementing approaches:** “[T]hose which try to implement some specific normative theory of ethics into autonomous agents so as to ensure that an agent acts in accordance with the principles of this theory” (Dyoub et al., 2020; p.7).
- **Bottom-up approaches,** by which “machines are expected to learn how to render ethical decisions through observation of human behavior in actual situations, without being taught any formal rules or being equipped with any particular moral philosophy” (Etzioni & Etzioni, 2017; pp.406-407)

References

- Awad, E., Levine, S., Anderson, M., Anderson, S. L., Conitzer, V., Crockett, M. J., ... & Tenenbaum, J. B. (2022). Computational ethics. *Trends in Cognitive Sciences*.
- Bonnemains, V., Saurel, C., & Tessier, C. (2018). Embedded ethics: some technical and ethical challenges. *Ethics and Information Technology*, 20(1), 41-58.
- Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y., Kramer, M.: Moral decision making frameworks for artificial intelligence (2017). Available at: www.aaai.org
- Dyoub, A., Costantini, S., & Lisi, F. A. (2020). Logic programming and machine ethics. *arXiv preprint arXiv:2009.11186*.
- Etzioni, A., & Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *The Journal of Ethics*, 21(4), 403-418.
- Geisslinger, M., Poszler, F., Betz, J., Lütge, C., & Lienkamp, M. (2021). Autonomous driving ethics: From trolley problem to ethics of risk. *Philosophy & Technology*, 34(4), 1033-1055.
- Jacobs, N., & Hultgren, A. (2021). Why value sensitive design needs ethical commitments. *Ethics and information technology*, 23(1), 23-26.
- Meier, L. J., Hein, A., Diepold, K., & Buyx, A. (2022). Algorithms for Ethical Decision-Making in the Clinic: A Proof of Concept. *The American Journal of Bioethics*, 1-17.
- Segun, S. T. (2021). From machine ethics to computational ethics. *AI & SOCIETY*, 36(1), 263-276.
- Woodgate, J., & Ajmeri, N. (2022). Principles for Macro Ethics of Sociotechnical Systems: Taxonomy and Future Directions. *arXiv preprint arXiv:2208.12616*.



Questions

A. Personal experience with computational ethics

1. Could you tell me about your professional background and your link with the topic we want to address today: formalizing ethical principles within algorithms/technology?
2. Overall, what is your opinion about formalizing ethical principles within algorithms/technology?

B. Reasons for & against computational ethics

3. What are reasons *for* formalizing ethical principles within algorithms/technologies? What *good consequences* would follow?
4. What are reasons *against* formalizing ethical principles within algorithms/technologies? What *negative consequences* would follow?

C. Implementing computational ethics

5. How can we implement computational ethics?
6. What exactly could and should we formalize within algorithms/technology? What could be underlying reference points for computational ethics?
7. What limits, if any, should there be to formalizing ethical principles within algorithms/technology?
8. What are important considerations/recommendations (for companies) when formalizing ethical principles within algorithms/technology?
9. What are necessary next steps to advance the field of computational ethics?
10. What are barriers to (companies) implementing computational ethics?
11. What, if any, are (fruitful) alternatives or complementary activities to formalizing ethical principles within algorithms/technology?

D. Closing questions

12. Are there any further questions that you think would be relevant in this context?
13. Do you know further experts that we may contact for our interview?

Thank you for your participation!

Appendix B of Essay I – Frequency analyses of the experts' arguments & recommendations

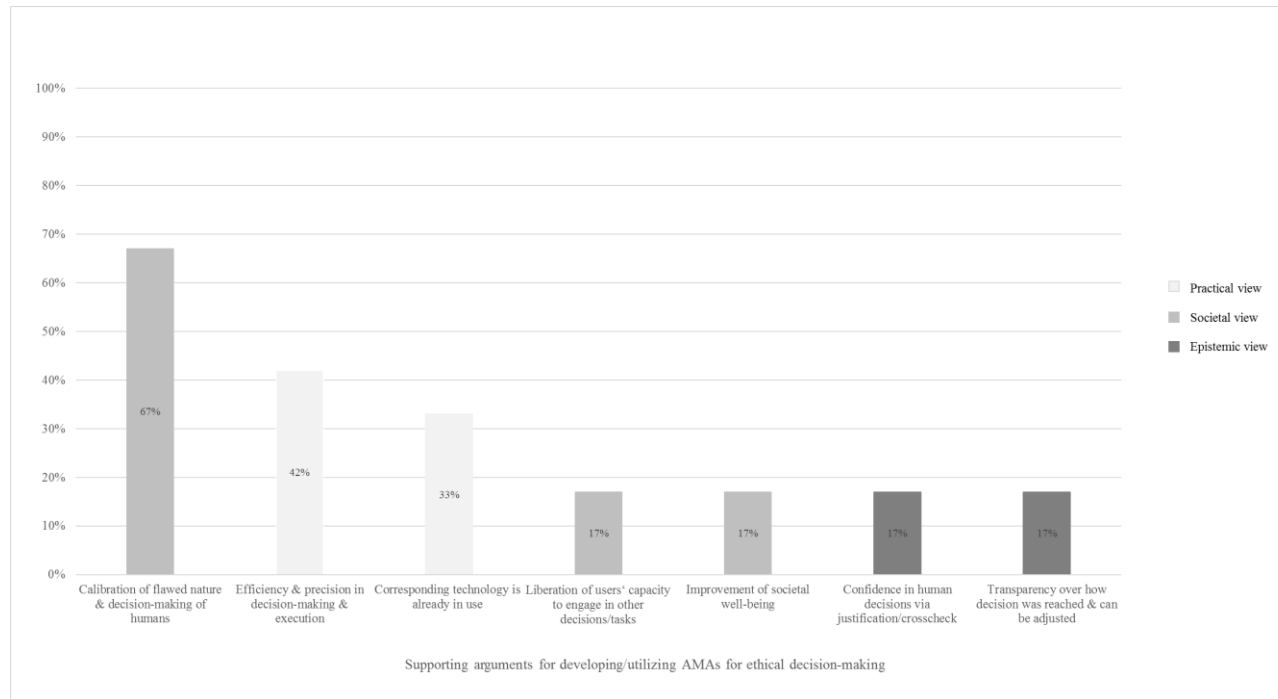


Figure 10: Experts' indicated arguments from a practical, societal, and epistemic view for developing/utilizing AMAs for ethical decision-making (ordered by frequency)

96 In total, the interviewed experts mentioned 7 different arguments supporting the development and use of AMAs for ethical decision-making, as illustrated in Figure 10. Out of these 7 arguments, 3 are related to societal aspects, 2 fall into the category of practical considerations, and 2 are of an epistemic nature. The argument that was mentioned most frequently (67%) relates to the societal view and stresses that AMAs help to calibrate the flawed nature and decision-making of humans. The second most frequent argument that was made by 42% of the interviewed experts relates to the practical view and addresses the efficiency and precision of AMAs in making (ethical) decisions. Benefits from an epistemic perspective were each only indicated by 17% of the interviewed experts, such as

Essay I: Formalizing Ethical Principles Within AI Systems: Experts' Opinions on Why (Not) And How to do it

installing confidence in users' decisions by utilizing the AMA as a crosscheck. Consequently, these arguments constitute the least frequently mentioned reasons supporting the development and use of AMAs for ethical decision-making.

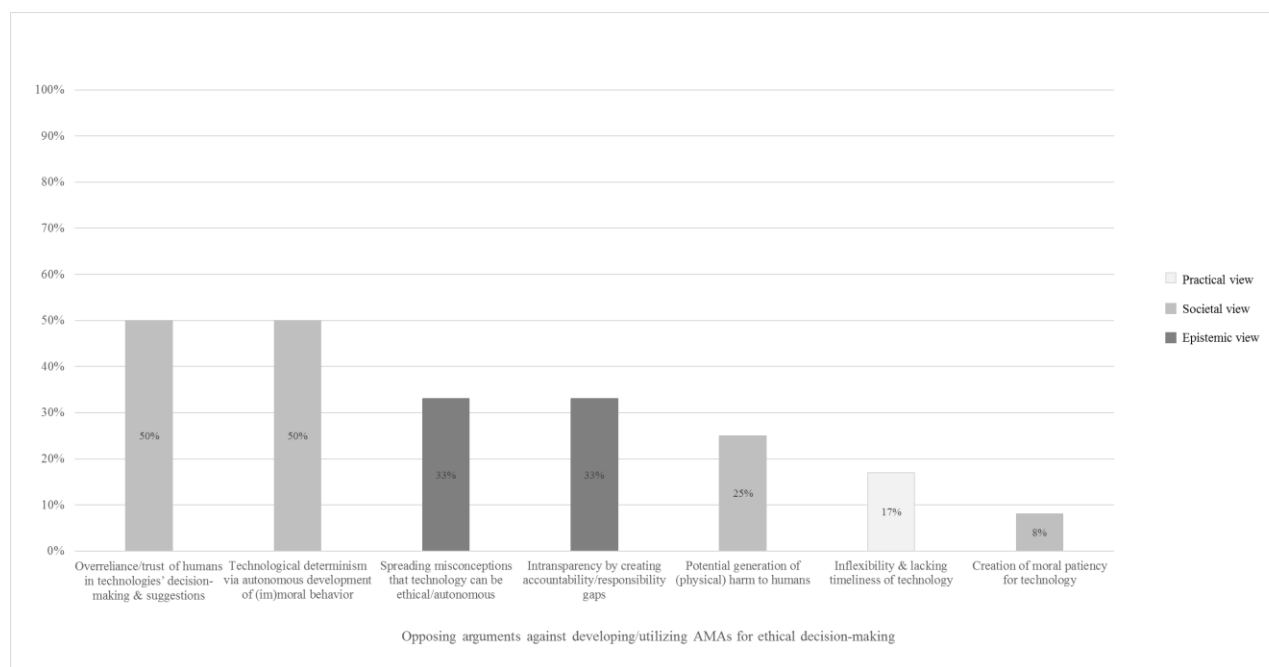


Figure 11: Experts' indicated arguments from a practical, societal, and epistemic view against developing/utilizing AMAs for ethical decision-making (ordered by frequency)

The interviewed experts mentioned 7 different arguments opposing the development and use of AMAs for ethical decision-making, as illustrated in Figure 11. Out of these 7 arguments, 4 are related to societal aspects, 2 are of an epistemic nature, and 1 falls into the category of practical considerations. The 2 arguments that were mentioned most frequently (50% each) relate to the societal view and highlight the overreliance of humans on AMAs as well as a potentially resulting technological determinism. After these arguments, the next 2 most frequently indicated reasons (33% each) against AMAs relate to the epistemic view, namely, these are: spreading the misconception that technology can be ethical/autonomous as well as the resulting intransparency or gaps in terms of

accountability/responsibility. The only argument from a practical view against AMAs concerns the inflexibility and lacking timeliness of technology, which was mentioned by 17% of the experts. The argument that was mentioned the least often with 8% relates to the epistemic view and stresses the creation of moral patency for pertinent technology.

Summarizing (as derived from Figure 10 and Figure 11), when discussing the benefits and drawbacks of AMAs for ethical decision-making, most arguments (7 in total) come from a societal perspective, while 4 arguments relate to an epistemic perspective and only 3 arguments are practical. The societal view not only dominates in terms of the total number of related arguments but also in terms of how frequently each argument of this perspective was indicated by the interviewed experts on average. Namely, the individual arguments relating to the societal view are mentioned by 33% of the experts on average, while the arguments relating to the practical view are mentioned by 31%, and arguments relating to the epistemic view are mentioned by 25% of experts on average.

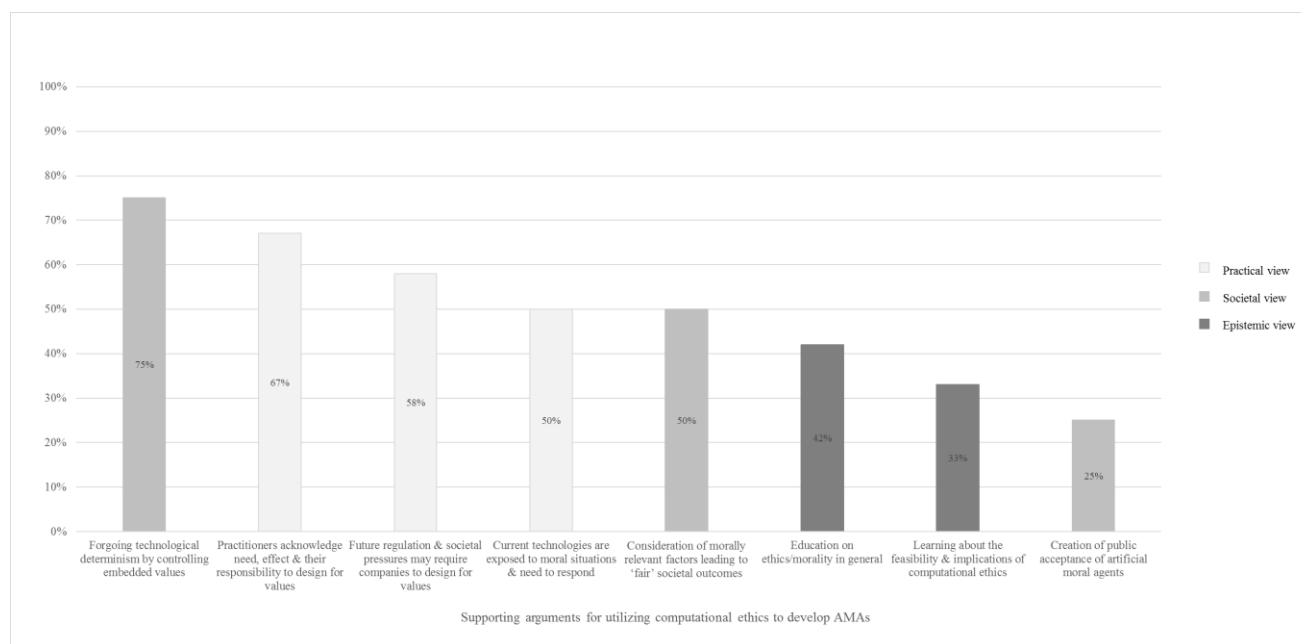


Figure 12: Experts' indicated arguments from a practical, societal, and epistemic view for utilizing computational ethics to develop AMAs (ordered by frequency)

As illustrated in Figure 12, the interviewed experts brought forward 8 different arguments supporting the use of computational ethics for developing AMAs, out of which 3 are practical, 3 relate to societal aspects, and 2 are epistemic considerations. The argument that was mentioned most frequently (75%) relates to the societal view and highlights that computational ethics grants control over what values are embedded and, thereby, acts as a measure against technological determinism. The next 2 arguments that were each indicated by the majority of experts relate to the practical view, namely, these are: practitioners acknowledge the need, effect, and responsibility to design for values (67%) as well as that future regulation and societal pressures may require companies to design for values (58%). Lower-tier arguments stem from the epistemic view, constituting 42% and 33%, respectively. The least frequently discussed argument (25%) pertains to the creation of public acceptance of AMAs, which can result when utilizing computational ethics to develop the respective systems.

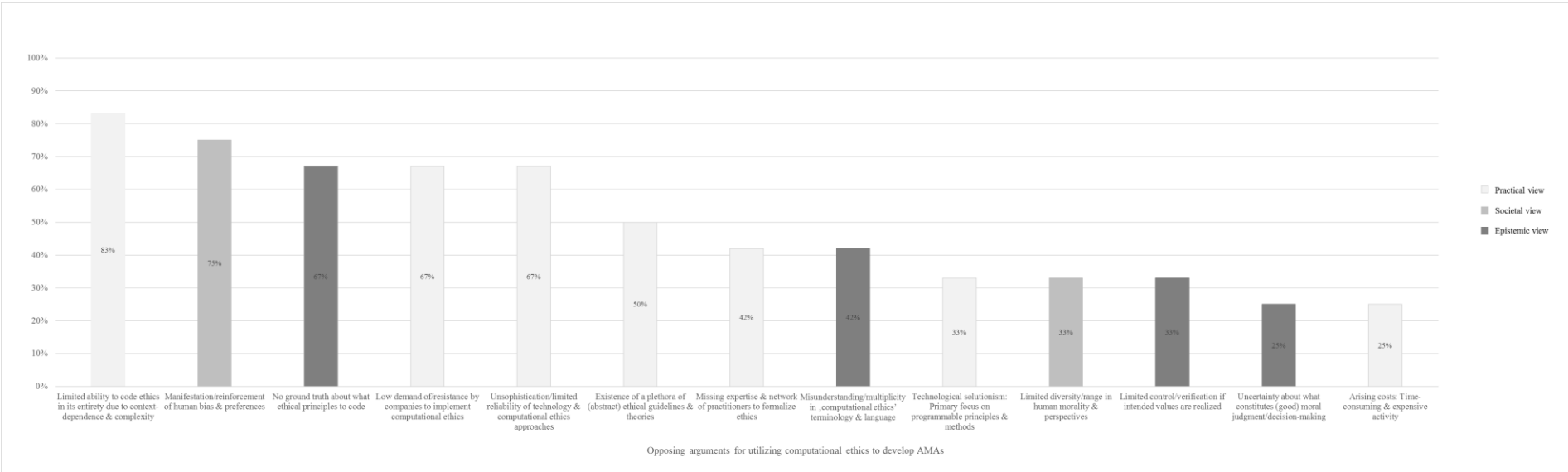


Figure 13: Experts’ indicated arguments from a practical, societal, and epistemic view against utilizing computational ethics to develop AMAs (ordered by frequency)

Figure 13 illustrates that the interviewed experts indicated 13 different arguments opposing the use of computational ethics for developing AMAs, out of which 7 are of a practical nature, 4 are epistemic considerations, and 2 relate to societal aspects. The argument that was mentioned most frequently (83%) relates to the practical view and highlights the limited ability to code ethics in its entirety. The second most frequent argument (75%) pertains to the societal view and emphasizes the manifestation or reinforcement of human biases and preferences that can result from formalizing ethical principles within AI systems. With a frequency of 67% each, three different arguments from the epistemic and the practical perspective were indicated, these are: there is no ground truth about what ethical principles to code, the low demand of/resistance by companies to implement computational ethics, as well as the unsophistication/limited reliability of technology and computational ethics approaches. The least frequently discussed arguments (25% each) relate to the uncertainty about what constitutes (good) moral judgment/decision-making as well as the extra costs that arise from executing computational ethics.

Summarizing (as derived from Figure 12 and Figure 13), when discussing the benefits and drawbacks of utilizing computational ethics to develop AMAs, most arguments (10 in total) come from a practical perspective, while 6 arguments relate to an epistemic perspective and 5 arguments relate to societal aspects. The practical view not only dominates in terms of the total number of related arguments but also in terms of how frequently each argument of this perspective was indicated by the interviewed experts on average. Namely, the individual arguments relating to the practical view were mentioned by 54% of the experts on average, while the arguments relating to the societal view were mentioned by 52%, and arguments relating to the epistemic view were mentioned by 40% of experts on average.

Essay I: Formalizing Ethical Principles Within AI Systems: Experts' Opinions on Why (Not) And How to do it

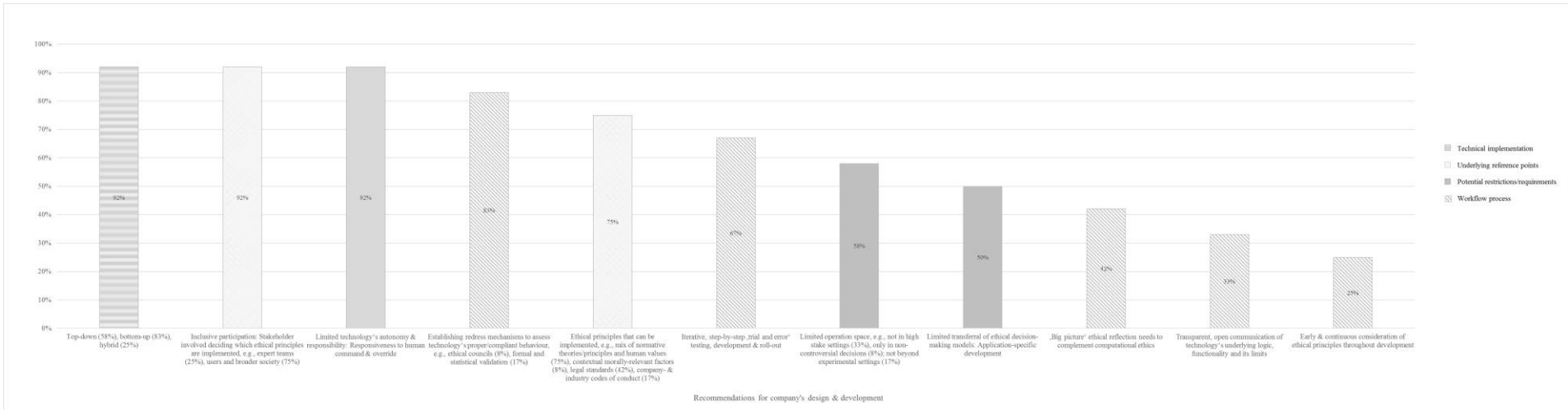


Figure 14: Experts' indicated recommendations for the company's design and development of AMAs (ordered by frequency)

As illustrated in Figure 14, the interviewed experts made 11 recommendations for companies when designing and developing AMAs. Out of these, 5 recommendations address how to structure the workflow process, 3 deal with potential restrictions or requirements to be taken into account, 2 involve underlying reference points for formalizing ethical principles, and 1 recommendation suggests methods for the technical implementation of computational ethics. Overall, 8 of the total 11 recommendations were indicated by at least half of the interviewed experts. The top recommendations that were each mentioned by 92% of the experts are threefold, namely, these are: various methods of how to technically implement computation ethics (e.g., in a top-down fashion), involving numerous stakeholders when deciding which ethical principles to implement as well as limiting the AMA's autonomy to ensure ensuring human control. 83% of the experts additionally suggested the establishment of redress mechanisms to assess the AMA's actual behavior. The least frequently mentioned recommendations pertain to the need to transparently communicate the technology's underlying logic (33%) and the early and continuous consideration of ethical principles throughout the AMA's development (25%).

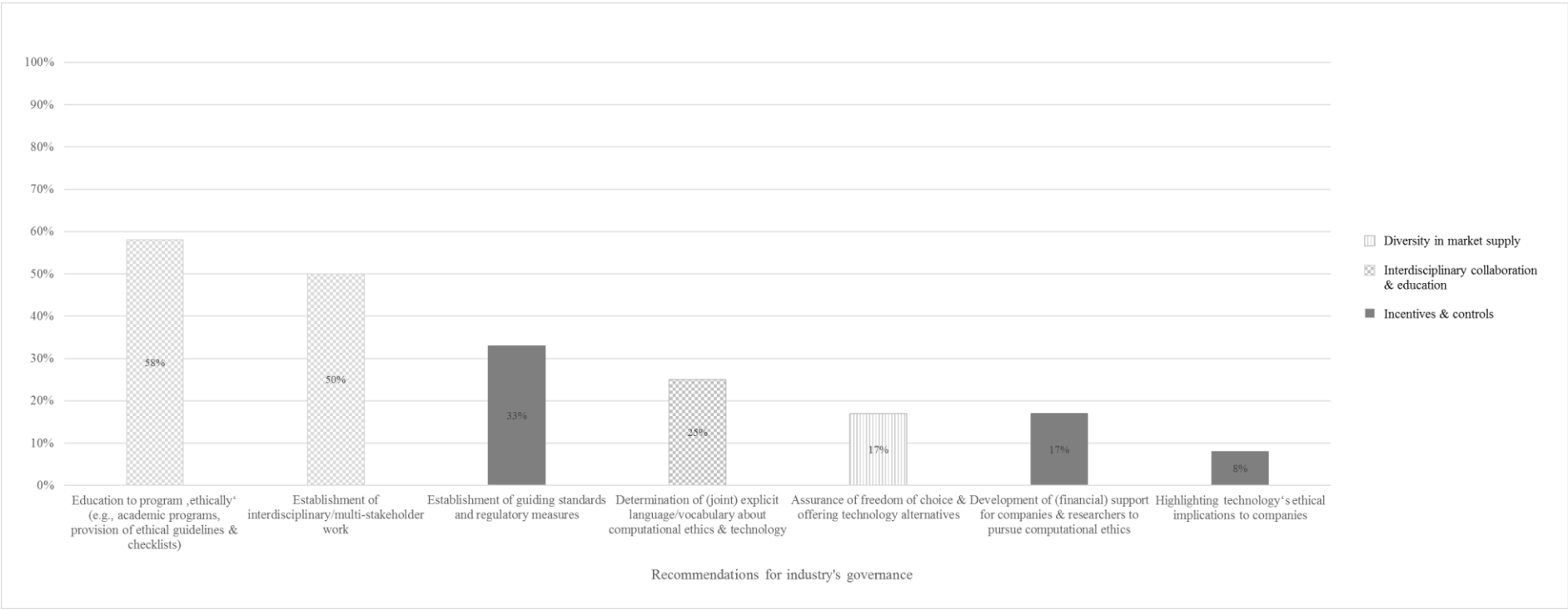


Figure 15: Experts’ indicated recommendations for the industry’s governance of AMAs (ordered by frequency)

Figure 15 illustrates the 7 recommendations that experts made in reference to the governance of AMAs and the corresponding industry. Out of these, 3 recommendations address interdisciplinary collaboration and education, the other 3 recommendations relate to the need for creating incentives and controls, while 1 recommendation links to diversity in the market supply. The two most frequent recommendations pertain to interdisciplinarity, namely: At least half of the experts mentioned the importance of interdisciplinary teams, projects, or education. Except for these two recommendations, at most, one-third of the interviewed experts endorsed any of the other recommendations. 33% of experts suggested the establishment of guiding standards and regulatory measures. Lower-tier recommendations entail the assurance of freedom of choice when it comes to consuming AMAs and the provision of (financial) support for companies

and researchers, with a frequency of 17% each. Lastly, 8% of the experts recommended that policymakers emphasize the ethical implications of AMAs to companies.

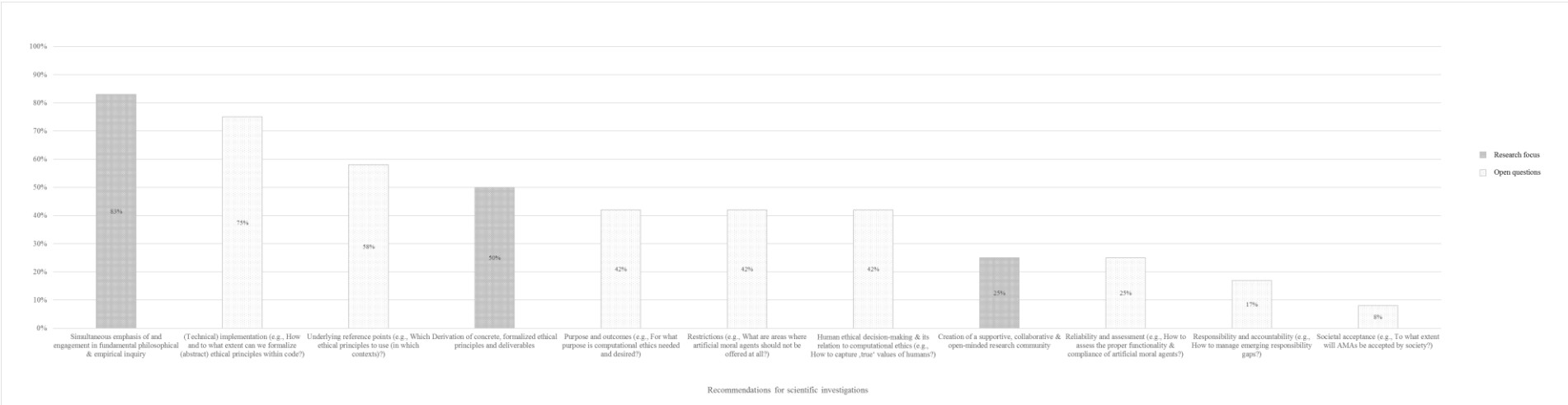


Figure 16: Experts’ indicated recommendations for the scientific community investigating computational ethics and AMAs (ordered by frequency)

The interviewed experts mentioned 11 different recommendations for the scientific community, as illustrated in Figure 16. Out of these, 8 recommendations emphasize open questions that should be examined in future research, and 3 recommendations relate to the overall structuring of the research ecosystem and activities. The recommendation that was mentioned most frequently (83%) pertains to the latter category and suggests the simultaneous engagement in fundamental philosophical and empirical research. The second most common suggestion underscores the requirement for additional exploration into the technical implementation and formalization of ethical principles, as highlighted by 75% of the experts. 58% of the interviewed experts called for further investigations of appropriate reference points that underlie AMAs. Furthermore, half of the experts emphasized that the research community is advised to provide more concrete and formalized deliverables in the future. Lower-tier recommendations relate to addressing open questions such as how to manage emerging responsibility gaps (as indicated by 17% of the experts) or determining the extent to which society will accept AMAs (as indicated by 8% of the experts).

3 | Essay II: Applying Ethical Theories to the Decision-Making of Self-Driving Vehicles: A Systematic Review and Integration of the Literature

Abstract

Self-driving vehicles will need to make decisions that carry ethical dimensions, and manufacturers have (a responsibility) to predetermine this underlying, deliberate decision-making process. With the rise of self-driving vehicles, scholars have simultaneously started investigating what ethical theories should guide machine behavior but have not concluded which ones should be preferred and adopted. We aim to address this matter by providing a holistic and analytical review of the autonomous driving ethics literature. Based on this review, we summarize the social, moral/legal, and functional advantages and disadvantages of applying particular ethical theories to the decision-making of self-driving vehicles. Furthermore, we derive a model that shows how the identified ethical theories could be technically implemented and integrated to guide the decision-making of self-driving vehicles. Overall, this article aims to lay the groundwork for a reflected, successful integration of ethical behavior within self-driving vehicles and deduce an updated research agenda.

Keywords: *self-driving vehicles, autonomous driving, ethical decision-making, machine ethics, philosophy*

Note: This chapter is based on a published paper co-authored by Maximilian Geisslinger, Johannes Betz, and Christoph Lütge. Therefore, the plural instead of the singular is used throughout this chapter. Author contributions to this paper and copyright information are summarized in Appendix B: Reference & copyright information by the publisher for the second essay (Essay II, Chapter 3) and Appendix D: Author contributions to the three essays in this dissertation.

Current publication status:

Poszler, F., Geisslinger, M., Betz, J., & Lütge, C. (2023). Applying ethical theories to the decision-making of self-driving vehicles: A systematic review and integration of the literature. *Technology in Society*, 75, 102350. <https://doi.org/10.1016/j.techsoc.2023.102350>

Link to the Creative Commons user license: <https://creativecommons.org/licenses/by/4.0/>

Conference presentations (of previous versions):

Poszler, F., Geißlinger, M., Betz, J., & Lütge, C. (July, 2022). Risk management at the center of ethics: The applicability of traditional ethical theories and how to truly advance the development of an ethical vehicle. Presented at The International Society of Business Economics and Ethics (ISBEE) World Congress, Bilbao.

Poszler, F. (May, 2022). ANDRE – AutoNomous DRiving Ethics. Presented at „Humanity in the Age of Intelligent Machines”, Leverhulme Centre for the Future of Intelligence, Cambridge. Available at: https://www.youtube.com/watch?v=UUjWlpPs_PM

3.1 Introduction

Self-driving vehicles (SDVs) – above the SAE automation level 4 – are designed to operate without human intervention (SAE International, 2018). Due to the expected advantages attributed to SDVs, such as higher comfort and safety for passengers, the industry is pushing ahead with the research, development, and testing of SDVs (Korosec, 2019). However, the introduction of SDVs onto the streets has already caused some accidents with detrimental consequences and fatalities (State of California – Department of Motor Vehicles, 2019). This implies that SDVs can end up in situations that entail life-and-death decisions and thus sometimes even have to make the highest ethical decision there is to make (Nyholm, 2018). Apart from these relatively rare dilemma situations, also in mundane traffic situations, SDVs will have to make decisions about how to distribute risks among traffic participants (Geisslinger et al., 2023), which similarly represents a decision of normative significance (Dietrich & Weisswange, 2019). Thus, whenever an SDV selects among different courses of action (i.e., trajectories), it engages in a kind of ethical decision-making process (Wallach & Allen, 2009) that needs to be preprogrammed by the technology's designers (Liu & Liu, 2021). Therefore, SDVs are not merely means to ends (i.e., a way for individuals to get from one place to another) nor value-neutral. Instead, they are bearers of morality (Latour & Venn, 2002), not only because they mediate safety and fairness aspects for individuals in traffic at all but because they do so bounded by the visions of the designers that are (implicitly) inscribed into technology (Akrich, 1992).

Having in mind the increasing autonomy and inherent morality of technologies such as SDVs, the field of machine ethics has emerged to explicitly direct the behavior of a particular technology towards 'appropriate' (as opposed to implicit/unreflected) norms (Nath & Salu, 2020). This can be achieved by equipping technologies (e.g., SDVs) with algorithms that follow an ideal ethical principle or set of principles (Anderson & Anderson, 2007) and, thereby, enable them to make use of explicit ethical considerations when deciding what to do in certain situations (Nyholm, 2023). Next to scholars, policymakers have raised the need to consider ethical dimensions in the programming of SDVs (e.g., European Commission, 2021; Kriebitz et al., 2022). In the future, companies may even be obligated to make their ethical programming explicit and transparent by, for example, disclosing their integrated ethical principles in a value register/list (e.g., compare ISO/IEC/IEEE24748-7000). In addition to potential legal obligations, it is assumed that explicitly considering ethical principles within the programming of SDVs, "i.e., making navigation decisions that are justifiable and reasonable", will lead to stronger public trust (Pickering & D'Souza, 2023; p.1) and will ensure wide adoption of SDVs (Etienne, 2022). Therefore, from an ethical, legal, and consumer perspective, it seems to be obsolete for automotive companies to neglect the ethical dimension of SDVs' decision-making and to program

fitting ethical principles. To identify ‘ideal’ ethical principles, it is not necessary to reinvent the wheel, but they can be drawn from existing forms of ethics (Sætra & Danaher, 2022). For example, they can be derived from philosophical theories to “guide decision-makers [e.g., SDVs] in making normative judgments and determining the moral permissibility of concrete courses of actions” (Woodgate & Ajmeri, 2022; p.1825). However, given the plethora of ethical theories (Segun, 2021b), it is questionable which one(s) to adhere to (Dyoub et al., 2020). Especially within the ethics of SDVs, there is an ongoing debate about which ethical theory “does the best job – and whether we have to make a choice, or whether we can combine insights from different theories” (Nyholm, 2023; p.70).

Correspondingly, investigations of ethical theories that should guide machine behavior in the field of autonomous driving have increased. For example, past literature has centered on solving a variety of modifications of Thomson’s (1984) ‘Trolley Problem’ and aimed to apply traditional normative theories such as utilitarianism, deontology, and virtue ethics to the decision-making process of SDVs (e.g., Etzioni & Etzioni, 2017; Gerdes & Thornton, 2015). Furthermore, scholars introduced other theories, such as risk ethics (e.g., Goodall, 2016a) or developed sector-specific ethical guidelines (e.g., Lütge, 2017; Lütge et al., 2021). This abundance of discussed options leaves academics, practitioners, and policymakers with a vast and unstructured literature providing no clear answer to the question of what constitutes ‘ethical’ decision-making of SDVs. In particular, researchers and practitioners continue to be divided about which (if any) ethical theories to advocate and about what implications (i.e., advantages or disadvantages) result from the integration of certain ethical theories into SDV. “[T]o identify the best sources of ethical theory that could help us to deal with this part of moral practice” (Nyholm & Smids, 2016; p.1277), the value-sensitive design approach suggests, amongst others, to carefully and comprehensively analyze the values and ethical theories that are to be implemented into certain technologies (Verbeek, 2011). This assessment can include the analysis of the theories’ moral justifications, their compliance with the law, their practical feasibility (Etienne, 2022), their compatibility with societal expectations (Krügel & Uhl, 2022), or their explainability (Németh, 2023).

This article aims to approach such an assessment and, thereby, help understand the best sources of ethical theory to program SDVs’ decision-making by systematically establishing a holistic overview and review of the various ethical theories that have been introduced in this regard in past literature. Therefore, this article differs from existing literature reviews in the field of autonomous driving ethics in the following ways. First, compared to previous literature reviews that addressed technical matters by, for example, summarizing the state of the art of SDVs (e.g., Deemantha & Hettige, 2022) or the impact of deep learning methods on vehicle behavior (e.g.,

Duncan, 2022), this article focuses on ethical considerations. Second, compared to previous literature reviews that summarized ethical considerations that relate to SDVs overall, such as safety, accountability, environmental sustainability, or data privacy issues (e.g., Hansson et al., 2021; Martinho et al., 2021; Soh & Martens, 2022), this article specifically focuses on the ethical decision-making of SDVs. After all, throughout existing reviews, scholars mentioned ‘moral safety choices’ (Papadimitriou et al., 2022), ‘algorithmic fairness’ (Holstein et al., 2021), ‘moral decision-making’ (Bergmann, 2022), and programming of SDVs in crashes (Nyholm, 2018) as one key issue. Thus, this is the first review zooming in on this particular issue by conducting a structured analysis of underlying ethical theories that could guide SDVs’ ethical decision-making. To this end, this article focuses on autonomous driving ethics literature and synthesizes its past publications to answer the following research questions:

1. *What are the advantages and disadvantages of applying particular ethical theories to the decision-making of self-driving vehicles?*
2. *How can ethical theories be integrated into the ethical decision-making of self-driving vehicles?*

To answer these questions, this paper is structured as follows. After having introduced the relevance of this research topic, details of the utilized methodology (i.e., a structured literature review) are elaborated. Then, key findings of the literature review are depicted, which constitute a summary of identified ethical theories as well as their respective social, moral/legal, and functional advantages and disadvantages when applying them to the decision-making of SDVs. Furthermore, combined theories, additional considerations/principles, and elaborate decision processes for the ethical decision-making of SDVs that have been sketched in past literature are highlighted. Afterward, the discussion section illustrates key takeaways and implications that can be drawn from past literature, provides a model summarizing how to integrate and apply ethical theories to SDVs’ decision-making, points towards caveats of this study, and opens questions to be investigated. Lastly, a short conclusion is drawn.

Altogether, the findings of this article provide guidance into when, why, and how particular ethical theories can be applied to the decision-making of SDVs, set the groundwork for the successful integration of ethical behavior into SDVs in the future, and deduce an updated research agenda.

3.2 Review method

We used a three-stage process to conduct an exhaustive review of the literature on autonomous driving ethics: (1) identification of relevant literature, (2) structural analysis of the literature, and (3) theme identification and article integration to synthesize the research (adopted from Theurer et al., 2018).

3.2.1 Stage I: Identifying relevant literature

Our comprehensive search approach is based on Webster and Watson (2002). The databases ScienceDirect, EBSCO Business Source Premier, Scopus, and Web of Science, as well as a forward-backward search, were utilized to ensure that all relevant journals and conference publications were considered in the investigation. To identify all existing literature, we did not limit our search to a specific date and included all publications until July 2023. This process resulted in 101 journal articles, book chapters, and contributions to conference proceedings identified as relevant and subjected to further analysis. A detailed description of our search processes, such as the date of collection, the applied search terms, or the entire search funnel, is provided in Table 7 and Figure 19 in Appendix A of Essay II – Literature search & analysis.

Inclusion and exclusion criteria

The formal inclusion criteria in the search process included the English language and primary and secondary studies. To enhance the review and provide a broad view of the topic, we did not constrain the search to peer-reviewed academic journal publications but included edited books and book chapters. Due to the novelty of the subject matter, we also considered conference proceedings and practitioner-oriented articles. Content-wise, publications that addressed the application of ethical theories to SDVs' decision-making processes were considered in scope. In contrast, excluded were publications that dealt with any topic other than this study's concrete focus. Namely, these 'out of scope' publications included those that did not remark on SDVs, ethical theories, and/or the ethical decision-making process of SDVs. In more detail, we excluded publications that focused on autonomous systems in general (i.e., with no specific link to SDVs) or those that investigated overall ethical and societal challenges arising from SDVs (e.g., economic and environmental issues). Furthermore, we did not include publications on human trust concerning SDVs or the governance of SDVs, in particular in terms of responsibility and liability. Lastly, we excluded publications that addressed ways of educating engineers to program 'ethically' or publications that discussed whether SDVs should make ethical decisions at all¹¹.

¹¹ In this review, we excluded the publications that solely debate the overall (moral) admissibility of SDVs to make ethical decisions. That is because this article is based on the premise that, in reality, SDVs will (be developed that) end

3.2.2 Stage II: Structural analysis of the literature

All 101 articles from stage I were entirely read and analyzed to provide a comprehensive basis for the subsequent identification of themes. The structural analysis addressed formal, exogenous article criteria and included the publication's

- issuing date (i.e., year)
- type (i.e., book chapter, conference proceeding, or journal article)
- discipline (i.e., Business, Computer Science, Engineering & Transportation, Ethics & Philosophy, Multidisciplinary Sciences, Social Sciences)
- utilized methodology (i.e., conceptual: theory, conceptual: mathematical, empirical: simulation; empirical: quantitative)
- degree of practicability (i.e., no remark on technical implementation, conceptualization of technical implementation, concrete suggestion of technical implementation)
- level of theory integration (i.e., no theory integration, theory integration)

The results of the structural literature analysis, including the chronological development of the literature, are provided in Appendix B of Essay II – Descriptive/structural analysis of the literature.

3.2.3 Stage III: Theme identification and article integration

In stage 3, an inductive coding and clustering process in an iterative and circular manner was conducted, meaning that no predetermined themes/codes were implied, but all themes/codes developed bottom-up from the datasets (i.e., the identified publications) themselves (Fereday & Muir-Cochrane, 2006). This entire content-based coding process was performed manually via the MAXQDA software. In line with Gioia et al. (2012), this coding process was structured into three steps: identification of 1st-order codes, organization of 1st-order codes into 2nd-order themes, and distillation of 2nd-order themes into aggregated dimensions. Based on our literature analysis, we identified three broad themes for categorizing the 'autonomous driving ethics' literature, namely:

1. identified ethical theories,
2. advantages & disadvantages of the identified theories, and
3. suggested solutions.

up in situations calling for ethical decisions. Thus, an investigation – as conducted in this article – of the suitability and applicability of various forms of ethics that could guide such decision-making processes is critical. Of course, the ultimate scope of implementation needs to be embedded in (meta)ethical considerations, as illustrated in Figure 18.

First, *identified guiding ethical theories* constitute applied/normative theories and ethical branches that were mentioned as underlying benchmarks to direct the ethical decision-making of SDVs in past literature, which are: deontology, virtue ethics, consequentialism/utilitarianism, contractualism, risk ethics, metaethics, and descriptive ethics. The second aggregated dimension entails stated *advantages and disadvantages* of applying each ethical theory to the functioning of SDVs. These advantages and disadvantages can be further clustered into social (i.e., society's inclination regarding a particular theory as well as its impact on society), moral/legal (i.e., a specific theory's congruity with human rights, responsibility, and liability) as well as functional (i.e., the feasibility and impact of transferring a particular theory into the software) arguments. Lastly, the dimension of *suggested solutions* comprises literature indications that have proposed combined frameworks (e.g., integrating traditional normative theories with risk ethics), addressed additional consideration/principles (e.g., the attention to situation-adjusted distributions) or sketched elaborate approaches of ethical decision-making processes for SDVs (e.g., data theories method). Figure 17 illustrates the corresponding data structure. The following findings section summarizes each dimension and the derived themes sequentially in more detail.

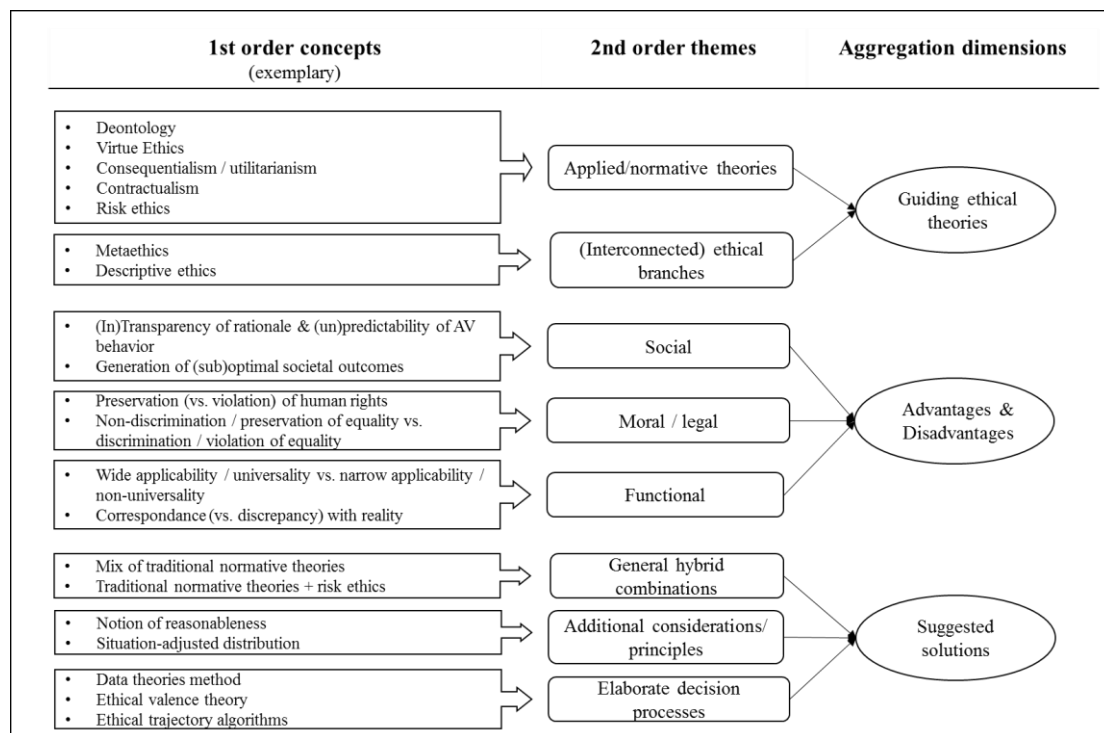


Figure 17: Data structure, reproduced from Corley and Gioia (2004)

3.3 Findings

The findings display the ethical theories that were identified in past literature in reference to the ethical decision-making of SDVs (3.3.1) as well as the stated advantages (3.3.2) and

disadvantages (3.3.3) of each theory in this regard. Furthermore, this finding section will highlight the first solutions of combined theories, additional considerations/principles, and elaborate decision processes (3.3.4) proposed in the respective literature.

3.3.1 Guiding ethical theories for the decision-making of SDVs

As illustrated in Table 3, scholars have highlighted guiding ethical theories¹² in past literature in the field of autonomous driving ethics, such as deontology, virtue ethics, consequentialism/utilitarianism, contractualism, risk ethics, metaethics, and descriptive ethics.

Guiding ethical theories	<i>Applied/normative theories</i>	<i>Deontology</i>	SDV decisions are determined based on their compliance with certain rules or constraints e.g., rules & constraints include: categorical imperative, Asimov's three laws of robotics, harm and collision avoidance, adherence to path/boundaries, doing nothing, compliance with traffic laws, sparing uninvolved participants
		<i>Virtue Ethics</i>	SDV decisions are determined based on the achievement of certain virtues e.g., virtues include: fairness, respect for authority & laws, carefulness, efficacy, responsibility, and courage
		<i>Consequentialism/Utilitarianism</i>	SDV decisions are determined based on their optimization of certain consequences and utilities e.g., consequences & utilities include: safety, energy consumption, passenger comfort, mobility/achievement, physical damage
		<i>Contractualism</i>	SDV decisions are determined based on methods and principles that are expected to yield hypothetical consent e.g., methods and principles include: veil of ignorance/original position, maximin/leximin
		<i>Risk ethics</i>	SDV decisions are determined based on balancing, distributing, and restricting certain risks e.g., risk considerations include: risk analysis and balancing based on risk parameters (vulnerability and speed of traffic participants, impact angle) or risk thresholds
	<i>(Interconnected) ethical branches</i>	<i>Metaethics</i>	SDV decisions are embedded in and shaped by high-level considerations e.g., high-level considerations include: the nature of moral vehicles, absolutism of ethical algorithms, admissibility of moral heteronomy
		<i>Descriptive ethics (i.e., Trolley Problem)</i>	SDV decisions are determined based on constraints, utilities, principles, etc. that align with societal preferences (which have been indicated in empirical studies or thought experiments) e.g., preferences include: utilitarian solutions, sparing humans over animals, sparing young and uninvolved individuals

¹² Note: The clustering of the identified guiding ethical theories is adopted from the 'forms of ethics' by Sætra and Danaher (2022).

Table 3: Identified ethical theories in past literature that can guide SDVs' ethical decision-making

Deontology. “If ethics categorizes actions and norms as morally correct or wrong, one then speaks of normative or prescriptive ethics” (Bartneck et al., 2021; p.19). Regarding traditional *normative theories*, scholars in autonomous driving literature have discussed the application possibilities of deontological, virtue, and consequential ethics, as well as contractualism. According to *deontological ethics*, the ethical correctness of actions is assessed based on the inherent characteristics of the action itself as opposed to the consequences the action would entail (Bartneck et al., 2021). For example, such a feature could entail whether a choice or action conforms to a particular ethical norm or duty (Alexander & Moore, 2007). Literature in the field of autonomous driving recognizes various moral norms or rules that could direct the underlying logic of SDVs. Stated rules include Kant’s categorical imperative and his prohibition on using humans as a mere means to an end (Fournier, 2016), Asimov’s three laws of robotics (e.g., law one declaring that a robot may not injure a human being) (Trappl, 2015), the obligation of harm and collision avoidance (Thornton et al., 2017), the adherence to designed paths (Gerdes & Thornton, 2015) or to virtual boundaries (Pickering & D’Souza, 2023), sparing uninvolved traffic participants (Németh, 2023), simply doing nothing (Zhao & Li, 2021), compliance with traffic laws (Pan et al., 2016) and the commitment to assertive driving (Shea-Blymyer & Abbas, 2021). Such basic principles or maxims could be applied to the functioning of SDVs in the form of priority rules, conditionals, or constraints with a hierarchal order to bind the system behavior (Gerdes & Thornton, 2015).

Virtue Ethics. *Virtue ethics* tends to grasp morality as a question of character, meaning that certain virtues such as prudence or courage are central to attaining a well-lived life. Such virtues, in principle, cannot merely be taught but are acquired naturally through experience (Hursthouse et al., 2016). Similarly to the concept of a virtuous individual, past literature has identified the following as potential characteristics that would mark a virtuous SDV: fairness, respect for authority and laws, carefulness (i.e., safety), efficacy, responsibility, and courage (i.e., the willingness to take a reasonable risk) (e.g., Gerdes, 2020; Pan et al., 2016; Nyholm, 2018). Such virtues could be used as positive signals in the reinforcement learning of an SDV (Geisslinger et al., 2021). In addition, literature focusing on virtue ethics in the field of SDVs suggests the adjustment of weights for different vehicles depending on the type of role or obligation the vehicle is assigned to serve (Gerdes & Thornton, 2015). For example, an autonomous ambulance could display a smaller weight on the virtue of respect for traffic laws (compared to an autonomous taxi) due to its Hippocratic Oath mandating to emphasize safety, i.e., to take patients to the hospital as quickly as possible (Wang et al., 2022a).

Consequentialism/Utilitarianism. According to *consequentialism*, “the ethical correctness of an action or norm solely [is determined] based on their (foreseeable) consequences” (Bartneck et al., 2021; p.20). This theory has been applied to SDVs mainly in the form of utilitarianism that aims at maximizing the overall good (Wang et al., 2020), which in the case of SDVs is often regarded as the minimization of victim numbers in car crashes (Johnsen et al., 2018). Apart from fatality minimization, additional utilities¹³ that could be considered costs or benefits have been identified. These include the amount of energy consumption, passenger comfort, mobility/achievement, physical damage, and compliance with traffic rules (e.g., Sütfield et al., 2019; Wang et al., 2022a). Based on such utilities, programmers of SDVs “would need to design cost-functions-algorithms that assign and calculate the expected cost of various options, selecting the one with the lowest cost” (Lin, 2016; p.73). Therefore, utilitarianism provides a method to construct the ethical decision-making of SDVs into an optimization problem (Wang et al., 2020).

Contractualism. In a broad sense, *contractualism* denotes the view that morality is based on a contract or agreement (Ashford & Mulgan, 2018). This theory can manifest itself as a political theory in the social contract tradition (e.g., Rawls, 1971) or as an ethical theory that aims to determine ethical principles (e.g., Scanlon, 1998). While Rawls (1971) aims to endorse social arrangements that every reasonable individual – under the veil of ignorance/from an original position – would agree to, Scanlon (1998) rather seeks principles that no one could reasonably reject, similar to a hypothetical consent (Ashford & Mulgan, 2018). Both strands have been mentioned in the autonomous driving literature. For example, a Rawlsian SDV would calculate the survival probability of each player for different actions and select the action alternative that players would agree to from an original position (Leben, 2017). The pertinent decision is said to be based on the principle of ‘maximin,’ which aims to improve the situation of the worst-off individual (i.e., the person bearing the lowest survival probability in a crash) (e.g., Jenkins, 2016; Leben, 2017; Dogan et al., 2020). Extensions of this principle constitute ‘leximin’ that can reach a decision even if survival probabilities are equal in two or more action alternatives by comparing, for example, the second-lowest survival probabilities (Keeling, 2017).

Risk Ethics. Risk ethics was coined by Nida-Rümelin et al. (2012) and “deal[s] with the morally correct actions in situations of risk” (Geisslinger et al., 2021; p.1041). Based on analyzing and calculating pertinent risks (Goodall, 2016a), risks need to be managed and balanced (Sütfield et al., 2019). SDVs could adopt this approach by calculating expected figures (i.e., the product of collision probability and estimated harm in case of unavoidable accidents) for involved parties

¹³ In general, most utilitarians are monistic in the sense that they argue for a single fundamental value, such as well-being or happiness, to be considered (Mason, 2018).

and across alternatives and, based on this calculation, decide on how to balance/distribute risks (e.g., Goodall, 2016a; Schäffner, 2018; Millán-Blanquel et al., 2020). Balancing and distributing risks would, in this sense, mean deciding on “who is put at marginally more risk of being sacrificed” (Bonnefon et al., 2019; p.503). Essential parameters in this calculation include the level of vulnerability and speed of traffic participants or the impact angle under which a collision would occur (e.g., Mirnig & Meschtscherjakov, 2019; Geisslinger et al., 2021). These parameters would then factor into the risk assessment and management strategy of the SDV as a result of its behavior, e.g., by determining the lateral position of the vehicle on the street: For example, if the SDV positions itself away from a cyclist, reduces the risk posed to the cyclist (e.g., Bonnefon et al., 2019; Geisslinger et al., 2021). Additionally, risk thresholds can be integrated into the ethical evaluation of a situation and, thereby, guide the critical decision of the SDV (Sütfeld et al., 2019).

In addition to the mentioned applied/normative theories, past literature has referred to ethical branches such as metaethics and descriptive ethics in the field of autonomous driving ethics.

Metaethics. *Metaethics* is generally defined as “a branch of analytic philosophy that explores the status, foundations, and scope of moral values, properties, and words, [...] metaethics focuses on what morality itself is” (DeLapp, 2011; p.1). Metaethical questions raised in the autonomous driving literature address the nature of moral vehicles, the absolutism of ethical algorithms of SDVs, and the admissibility of moral heteronomy. Regarding the *nature of moral vehicles*, literature tried to answer the question: What makes an SDV moral? In this respect, the required and good virtues and objectives of SDVs mentioned in the literature include, for example, responsibility, fairness, or respect for authority (e.g., Pan et al., 2016). Furthermore, if fairness is one of these virtues, what does this mean in concrete terms (e.g., fairer, to whom?) (e.g., Siegel & Pappas, 2021)? In terms of studies on the *absolutism of ethical algorithms*, scholars have discussed whether a single philosophical framework (‘one-size-fits-all’ solution) is sufficient for programming SDVs or whether a plurality of moral truths is to be considered, especially when it comes to geographical and cultural differences (e.g., Reed et al., 2021). Generally, it is acknowledged that there is no clear proposition on whether there is or should be a preference for one specific framework (e.g., Karnouskos, 2018) or whether the one ‘right’ framework exists at all (e.g., Siegel & Pappas, 2021). Therefore, “[r]ather than searching for a single ethical principle that should guide the programming of SDVs – a so-called ‘prime directive’ – programming ethics should be considered as an engineering problem that allows a range of solutions” (Jenkins, 2016; p.6). This suggestion is directly linked to the next metaethical issue that was identified in the autonomous driving literature: the admissibility of moral heteronomy. Regarding the *admissibility of moral heteronomy*, researchers have investigated to what extent the decisions and actions taken by SDVs could or should be determined by actors other than the drivers themselves, such as the

SDV's programmers (e.g., Keeling et al., 2019) or the government (Gurney, 2016). For example, instead of a mandatory ethics setting that may threaten self-determination and enact paternalism, SDVs could be equipped with a personal ethics setting that enables the passenger to determine the distribution of risks in life or death decisions (Gogoll & Müller, 2017). Reed et al. (2021) describe an ethical goal function that "should be defined by governmental bodies with input from citizens and stakeholders" (p.777) to adapt to society's preferences dynamically. Similarly, Krügel and Uhl (2022) argue the programming of SDVs' ethical decision-making should be determined in a participatory manner by involving all those who will be affected by the technology.

Descriptive Ethics. *Descriptive ethics* describes and explains normative systems (Bartneck et al., 2021). In autonomous driving research, descriptive ethics has been adopted by assessing individuals' preferences regarding the functioning and decisions of SDVs, especially in moral dilemmas in the context of unavoidable accidents. In doing so, scholars have often drawn on thought experiments, such as the *Trolley Problem*, as a tool/methodology to assess societal reasoning and preferred ethical decisions in unavoidable accidents where tradeoffs/sacrifices between individuals need to be undertaken (e.g., Awad et al., 2018). Respective studies have discovered many differing preferences indicated by participants. First, the results of empirical studies seem to suggest a general preference for utilitarian solutions, in particular, the tendency to minimize harm, i.e., the number of victims (Awad et al., 2018; Faulhaber et al., 2019; Bonnefon et al., 2016). Moreover, participants expressed preferences to spare humans over animals (Awad et al., 2018), women over men (Sütfeld et al., 2017), young individuals (Bergmann et al., 2018), and uninvolved individuals such as pedestrians (on sidewalks) (Wintersberger et al., 2017) while they displayed a willingness to sacrifice norm violators (Frank et al., 2019). Other preferences that transpired (less universally and coherently) included sparing the rich over the poor (Awad et al., 2018) and individuals of high social status (Altay et al., 2023), favoring self-preservation (Meder et al., 2019), or, on the contrary, a limited willingness to commit self-sacrifice (Faulhaber, 2019) and the inclination to spare friends or family over strangers (Frison et al., 2016). Scholars have argued that the results of such studies could be implemented into SDVs by simply mimicking actual human moral opinions (e.g., Goodall, 2014a).

Overall, while metaethics are relevant as high-level considerations when programming SDVs (compare Figure 18), this branch of ethics does not intend to directly instruct the internal decision-making of SDVs compared to the applied/normative theories and insights from descriptive ethics. Thus, these latter types of ethical theories are the focus of study in the following two sections. As will be shown, these guiding ethical theories entail certain advantages and disadvantages when applied to the ethical decision-making of SDVs.

3.3.2 Advantages of applying particular ethical theories to the decision-making of SDVs

Past literature has highlighted several advantages of applying ethical theories to the decision-making of SDVs. A holistic overview of identified benefits in past literature is illustrated in Table 4¹⁴. The advantages can be clustered according to social, moral/legal, and functional arguments.

Social advantages. Social advantages include the transparency of rationale & predictability of SDV behavior, the generation of optimal societal outcomes, and compatibility with societal preferences/agreements. Concerning the *transparency of rationale & predictability of SDV behavior*, for most ethical theories, reasons for how they contribute to this advantage are listed. For example, programming an SDV in a deontological manner generates clarity over the SDV's reasoning and behavior for society due to its constant functioning on fixed priorities and constraints. Implementing a virtue ethics approach may offer predictability, assuming that consulted virtues entail stable patterns of SDV actions. Contractualism, instantiated in the form of the Maximin principle, features a transparent decision procedure over numerical figures that are known to operate to the advantage of the worst-off traffic participant. Therefore, whatever the circumstances will be, individuals can expect a result that bypasses the worst outcome. Risk ethics approaches are transparent in that their cost function can be inspected retrospectively to determine the SDV's logic and calculation that led to a particular decision. More generally, empirical research dealing with individuals' attitudes or thought experiments such as the Trolley Problem can familiarize society with different ethical theories. This creates transparency about the implications of integrating particular ideas into SDVs in the first place. Concerning the *generation of optimal societal outcomes*, only three of the considered theories generate pertinent advantages in past literature. For example, implementing a consequentialist/utilitarian calculus into SDVs would produce the lowest number of fatalities or could achieve Pareto-efficient outcomes for society when connected with the restriction that it is only justified to harm someone if they would receive at least the same harm on any alternative. SDVs operating on the Maximin principle can be considered beneficial for society overall in that they protect the most vulnerable traffic participants. Applying risk ethics approaches to the functioning of SDVs is necessary to mitigate collisions in the first place and to avert that specific "evil" (by accounting not only for the severity but also for the probability of the outcome). Therefore, risk ethics approaches contribute to the generation of optimal societal outcomes.

For all identified theories, indications are made that support their *compatibility with societal preferences/agreements*, meaning that the adoption of a particular theory corresponds to how

¹⁴ The following text passages summarize overall identified advantages without stating specific references. All corresponding references for each advantage are indicated in Table 4.

individuals (would like SDVs to) make decisions. For example, a study has indicated that deontological ethics, when adopted in the form of clear-cut rules, are well-perceived and accepted by society. Virtue ethics may similarly meet society's approval as it considers special obligations and takes into account the role morality. It does so by emphasizing differing virtues depending on the particular SDV type (e.g., private SDVs may be obligated to assure passenger safety, or for self-driving ambulances, it may be forbidden to harm third parties, while they may be allowed to neglect traffic laws). SDV decisions that follow a utilitarian calculus (i.e., minimizing the total fatalities) also seem to align with individual preferences, as revealed by a few empirical studies in the past. Contractualism may be practicable in achieving societal agreement in the field of autonomous driving ethics by offering a process (i.e., the veil of ignorance) that helps to determine what constitutes a generally accepted harm distribution that is executed by SDVs. In addition, according to a previous empirical study, participants accepted decisions that aligned with the Maximin principle. Furthermore, risk ethics approaches allow the consideration of risks in a decision situation, which plays a significant role in actual human decision-making. Therefore, considering risks is key to establishing an SDV decision that resembles societal decision-making processes. Using the results from empirical investigations would generate insights into individuals' preferences that could then be utilized for the ethical theories or values embedded in the decision-making of an SDV. Thus, descriptive ethics are a first step toward the possibility of conforming to individuals' preferences and mimicking individuals' (ethical) driving behavior, which, in turn, can increase public trust in the technology.

Moral/legal advantages. Moral/legal advantages include preservation of human rights, non-discrimination/preservation of equality, as well as consideration of obligations & responsibility. Concerning the *preservation of human rights*, for five identified theories, arguments are stated in past literature that illustrate how they may meet this benefit. For example, SDVs following deontological ethics and Rawlsian algorithms may implicate respect for personal rights in that individuals and their interests are not treated as instrumental values that can be set off against each other or used to increase the collective good at their expense. However, the utilitarian calculus may not entail a violation of human dignity and personal rights if, in the form of general programming aimed at minimizing overall harm, where involved and affected parties are unknown. In addition, consequentialist approaches correspond to legal standards currently being developed. Risk ethics approaches can be argued to preserve human rights because they do not directly implicate the offset of individuals' lives but rather a decision of who is put at marginally more risk of being targeted. Managing dilemma situations by principles of risk distributions (amongst others) has been set out in recent regulatory drafts, and thus, risk ethics approaches achieve legitimacy from a legal perspective. Concerning descriptive ethics, indicated preferences

within well-established empirical research may bear normative value to legitimize a particular SDV's functionality.

Concerning the *non-discrimination/preservation of equality*, pertinent arguments for three theories are found in past literature. Particular models of Kantian ethics may imply selecting those rules as universal laws that apply to all individuals equally without unjustified advantage toward particular parties. Also, the utilitarian calculus could satisfy the principle of non-discrimination since general programming that reduces the total number of personal injuries will distribute ex-ante risks equally among all traffic participants. Similarly, contractualism – executed in the form of the veil of ignorance – will remove the possibility of self-interested decisions so that mutually advantageous and impartial decisions for all traffic participants derive.

Concerning the *consideration of obligations & responsibility*, pertinent remarks are made for only two of the selected theories. On the one hand, virtue ethics allows for moderation in an SDV's behavior based on its particular role or vehicle type. On the other hand, risk ethics approaches will enable the consideration of obligations and responsibility by integrating a compensation scheme by shifting risk allocations away from innocent parties toward parties that bring risks into traffic in the first place.

Functional advantages. Functional advantages include broad applicability/universality, compatibility & effectiveness of implementation, and correspondence with reality. Concerning the *broad applicability/universality*, all identified ethical theories offer corresponding benefits to some extent. Deontological ethics offers broad applicability since its general rules can provide guidance in many traffic situations. For example, the rule for an SDV to “stay in the lane or on a particular path” could be executed or imaged in most traffic scenarios. Furthermore, it can be argued that virtue ethics meet broad applicability, as it does not generate a one-fits-all solution; instead, different weights/virtues can be adapted for different types of vehicles. Using utilitarianism as the basis for programming SDV behavior would allow flexible reactions that can adjust to numerous traffic situations depending on which of the set utilities are present in the SDV's environment at a particular time. Contractualism – in the form of the Maximin principle – is widely applicable since it produces a unique choice of action (i.e., the one with the least bad worst outcome) in every situation where such an action is available. Applying risk ethics approaches to the decision-making of SDVs also leads to one recommended action, not just in dilemma situations but – by means of adjusting an SDV's lateral positioning – at every moment of operation. Moreover, risk ethics settings could enable up-to-date functionality. For example, probabilities can be revised according to recent experiences or crash statistics, while magnitudes can be adjusted to reflect societal preferences. Therefore, past literature seems to argue that risk ethics approaches achieve wide applicability in many traffic scenarios simultaneously and from a

temporal perspective. Moreover, traditional trolley cases may have stipulated away issues such as the existence of uncertainty, leaving it somewhat narrowed down to the theoretical sphere with no actual applicability for practice. However, in principle, thought experiments could be adaptable in that utilized scenarios can be changed and enriched with different factors to be investigated.

Concerning the *compatibility & effectiveness of implementation*, again, all selected theories are said to be associated with this benefit to some degree, which, amongst others, is illustrated by the fact that each theory has been computed for test purposes in at least one article. First, using deontological ethics in the programming of SDVs is technically feasible as machines can easily follow defined rules and constraints. Using machine learning techniques, SDVs can independently generate moral rules that are to be followed. Furthermore, drawing on deontological ethics in the programming of SDVs can be argued to be effective since, for its high-level regulations and decisions, less information (e.g., what the outcomes of alternative actions would be) is needed. Virtue ethics may be technically implementable by determining and utilizing particular desirable virtues as positive signals in the reinforcement learning of SDVs. Integrating utilitarianism, in the form of a cost function that calculates the outcomes of various scenarios and selects the one that maximizes utility, corresponds well to the functionality of SDVs as it is designed to optimize preset functions. Utilitarian calculations can also be conducted faster and more efficiently by SDVs than by human minds. The same argument can be made for the Maximin principle, which is implementable as a utility function and selects the action that maximizes the lowest payoff. The Maximin principle can also serve as a design criterion for rewards in reinforcement learning. Risk ethics approaches are similarly applicable by quantifying, comparing, and selecting among decision alternatives while not only considering utility or harm but also adding the factor of probability to its calculation. Risk ethics approaches are computable in the form of risk thresholds or classifications that are not to be exceeded. Descriptive ethics contribute to the effectiveness of implementation in that assessing and ultimately integrating the identified societal preferences into SDVs will spur acceptance and adoption levels of SDVs. Preferences indicated in empirical investigations can be decisive for, for example, which ethical theories are (to what extent) integrated within SDVs. Another possibility for implementing descriptive ethics would be to mimic human ethical decision-making through machine learning.

Concerning *correspondence with reality*, only risk ethics and descriptive ethics list corresponding arguments in past literature. In support of risk ethics approaches, scholars argue that driving scenarios, in reality, are characterized by an excellent level of dynamics, sources of uncertainty, and a variety of possible outcomes. Similarly, state-of-the-art SDV algorithms usually operate on probabilities and not on calculations of actual damage, which suggests that risk

ethics approaches highly coincide with practice. The use of indicated moral preferences in empirical studies could be more relevant and meaningful than ethical theories for an SDV's functioning and decision-making since individuals usually are not guided by deontic or utilitarian grounds but rather by normative standards of their culture. Furthermore, machine-learning approaches that mimic human behavior are said to reflect manufacturers' actual activities and industry standards (e.g., Tesla SDVs supposedly mimic human behavior) (Siegel & Pappas, 2021).

Essay II: Applying Ethical Theories to the Decision-Making of Self-Driving Vehicles: A Systematic Review and Integration of the Literature

Evaluation criteria for advantages		Guiding ethical theories					
		<i>Deontology</i>	<i>Virtue Ethics</i>	<i>Consequentialism/ utilitarianism</i>	<i>Contractualism</i>	<i>Risk ethics</i>	<i>Descriptive Ethics (Trolley Problem)</i>
social	Transparency of rationale & predictability of SDV behavior	<ul style="list-style-type: none"> Coherent, well-defined rules such as “Stay in lane and perform an emergency stop” are transparent, even under limited information about the situation at hand [25; 60; 64] Clarity and consistency of reasoning and the relationship between an SDV’s behavior and its programming due to fixed rules in retrospective assessment but also in real-time for human drivers [42; 64; 92] Sets clear priorities, goals and constraints [15; 21; 29; 33; 44; 76; 74; 92] 	<ul style="list-style-type: none"> Offers consistent and (retrospectively) explainable/transparent behavior, assuming virtues have developed into habits/are the result of a reasoning process [53; 81] 		<ul style="list-style-type: none"> Maximin features a clear decision procedure over numerical figures [54] 	<ul style="list-style-type: none"> SDV’s logic/risk calculation can be examined to determine why it behaved in a certain way [29; 30; 36] 	<ul style="list-style-type: none"> Raises understanding and awareness of the different ethical options to program SDVs [10; 13; 22; 25; 37; 41; 52; 63; 70; 72; 78; 80; 88; 91]
	Generation of optimal societal outcomes			<ul style="list-style-type: none"> Ensures the maximization of group interest and produces the lowest number of fatalities [18; 29; 30; 45; 53; 56; 77] If restricted, it corresponds to pareto-efficient outcomes in that a dominating outcome has strictly greater utility than the dominated outcome [49] 	<ul style="list-style-type: none"> Potentially corresponds to pareto-efficient outcomes in that individuals cannot reasonably/rationally reject the selected alternative [42; 48; 49; 54; 77] Protection of most vulnerable individuals in accident scenarios through Maximin [30; 54] In line with the law of diminishing marginal utility, giving additional benefits to the worst-off will increase the overall benefit [86] 	<ul style="list-style-type: none"> Severity of outcomes & frequency/probability considered so that the “greatest possible evil” can be averted [8; 36; 46; 52; 57; 86] Risk management will often be necessary to mitigate collisions in the first place [58; 91] 	
	Compatibility with societal preferences/ agreements	<ul style="list-style-type: none"> Coherent, well-defined rules have a strong positive impact on public opinion [60; 88] 	<ul style="list-style-type: none"> Acceptable due to consideration of special obligations, role morality/moderation, for example, an SDV’s obligation to protect its passengers [28; 29; 92] 	<ul style="list-style-type: none"> Given the uncertainty of one’s position, utilitarian optimization is the rational choice [6; 7; 42] (especially when it comes to public SDVs) [67] Past studies highlighted that individuals indicated preferences for “utilitarian SDVs” that minimize the total number of fatalities [6; 9; 14; 17; 24; 26; 27; 28; 29; 30; 39; 47; 55; 56; 63; 64; 66; 75; 80; 90; 98; 99; 101] 	<ul style="list-style-type: none"> Based on the self-interest of each person to maximize their own chances of surviving, allocations that are achieved through this process (e.g., behind the veil of ignorance) would be generally justifiable to, rational for, and preferred by the recipients of harm [12; 37; 42; 49; 54; 72] Respondents accepted decisions made by an algorithm that is based on the Maximin principle [77] 	<ul style="list-style-type: none"> Risk consideration demanded/ (implicitly) accepted, probably humans adopt an implicit form of risk reaction in their natural behavior as well [28; 29; 32; 52; 61; 97] Different decision-making under risk [41; 50; 64; 73; 80; 83; 84; 98] 	<ul style="list-style-type: none"> Based on insights from surveys, the potential to conform with people’s preferences, for example, concerning the preference of certain weights [4; 5; 19; 25; 29; 59; 80; 82; 101] Gives insight into societal reasoning, opinions, and evaluation of ethical theories & their relevance for the functionality of SDVs [10; 14; 25; 36; 38; 41; 44; 50; 57; 61; 66; 71; 72; 73; 78; 81; 83; 88; 91; 101] Human-like driving characteristics can improve trust in SDVs [70]
moral/ legal	Preservation of human rights	<ul style="list-style-type: none"> Corresponds to modern liberalism in that individuals’ rights are respected, i.e., they are not set off against another or instrumentalized for the greater good [6; 15; 26; 40; 47; 49] No interference with existing threat, i.e., no killing or intentional harm, 		<ul style="list-style-type: none"> General programming aimed at minimizing total harm seems justifiable and entails no violation of human dignity if it is a matter of probability prognosis in which identities are unknown [6; 21; 29; 31] 	<ul style="list-style-type: none"> Rawlsian algorithm implicates unwillingness to sacrifice the interests of one person for the interests of others [48; 54] 	<ul style="list-style-type: none"> No direct calculation/decision over life and death but a decision about who is put at marginally more risk of being sacrificed [7; 10; 29; 30] Corresponds with policymakers’ calls to consider risk management and balancing when SDVs face dilemmas [30] 	<ul style="list-style-type: none"> Preferences stated in well-conducted empirical research may have legitimate normative implications [19; 59]

Essay II: Applying Ethical Theories to the Decision-Making of Self-Driving Vehicles: A Systematic Review and Integration of the Literature

functional		which matters for liability attributions [33; 49; 54; 57; 78; 83]		<ul style="list-style-type: none"> Consequentialist approaches align with existing standards such as ISO 26262 [32] 			
	Non-discrimination/ preservation of equality	<ul style="list-style-type: none"> Kantian ethics may imply that all individuals are treated equally, with no unjustified advantage in crash scenarios [15; 25; 72] 		<ul style="list-style-type: none"> General programming to reduce the total number of personal injuries may be justifiable as (ex-ante) risks will be distributed equally among road users [7; 62] 	<ul style="list-style-type: none"> Veil of ignorance removes the possibility of self-interested decisions/unfair bargaining advantages so that mutually-advantageous principles are enacted [25; 48; 54] 		
	Consideration of obligations & responsibility		<ul style="list-style-type: none"> Ability to consider special obligations, role morality/moderation depending on the type of vehicle (e.g., publicly owned vs. private AV, ambulance) [28; 29; 53; 57; 74; 92] 			<ul style="list-style-type: none"> Allocation of risk can be used as a form of compensation, e.g., higher allocation of risk for those who actively bring risk into traffic [7; 15; 31; 42; 58] 	
	Wide applicability/ universality	<ul style="list-style-type: none"> Ability to provide guidance in many traffic scenarios [21; 95] 	<ul style="list-style-type: none"> Computable in the form of qualitative adjustments as weights to determine the strengths of applied rules and costs for different types of vehicles [32; 33; 53; 74; 92] 	<ul style="list-style-type: none"> Allows flexible reactions that can adapt to circumstances [29; 92] Many situational factors could be considered in its calculation [19; 29] Yields definite course of action, no standstill [6; 29; 45] 	<ul style="list-style-type: none"> Maximin produces a unique decision in almost every situation [54] 	<ul style="list-style-type: none"> Enables an implementation independent of the situation, thus not limited to dilemma or crash situations but also applicable in mundane traffic scenarios [29; 30] Risk ethics settings are easily adjustable when SDV is behaving unsafely or unexpectedly [36] Always recommends one action [36] 	<ul style="list-style-type: none"> Easy adaptable to check against and imagine new, differing scenarios [2; 4; 6; 25; 36; 42; 78]
	Compatibility & effectiveness of implementation	<ul style="list-style-type: none"> Corresponds to the functionality of machines which are designed to follow defined (hierarchies of) rules [19; 29; 33; 34; 68; 80; 81; 87; 92; 94; 96] Computable in the form of (soft) constraints on the SDV's behavior or to guide rewards/goals in reinforcement learning [25; 34; 69; 74; 76; 77; 78; 87; 92] Moral rules can be generated by machine learning [68] Requires less information, e.g., no considerations of alternatives or likelihoods necessary [64] Examples that computed deontology: [18; 23; 33; 53; 68; 69; 70; 74; 76; 87; 92; 94] 	<ul style="list-style-type: none"> Predetermined virtues can be utilized as positive signals in reinforcement learning [29] Examples that computed virtue ethics: [53] 	<ul style="list-style-type: none"> Corresponds to the functionality of machines which are designed to follow structured guidelines/to maximize the optimization of preset functions [29; 33; 35; 80; 92; 95; 96] Computable in the form of a mathematical description or cost function that calculates the outcomes of various scenarios and selects the optimal cost-benefit ratio (e.g., an option that minimizes the total number of fatalities) [5; 15; 25; 29; 33; 34; 48; 67; 78; 86; 92; 100] SDVs would be able to make utilitarian calculations of different options quicker and more reliable than humans [40; 66; 72] Examples that computed utilitarianism: [18; 20; 23; 29; 30; 33; 45; 53; 68; 69; 70; 87; 92; 94] 	<ul style="list-style-type: none"> Maximin is computable in the form of utility functions that are based on a set of action profiles and select the action that maximizes the lowest payoff [15; 54] Maximin principle can be used as design criteria for rewards in reinforcement learning [77] Examples that computed contractualism: [20; 23; 29; 30; 45; 77] 	<ul style="list-style-type: none"> Possibility to quantify & compare decision alternatives, e.g., by considering the cumulative risk of certain outcomes [29; 30; 36; 43; 51] Computable in the form of risk thresholds or classifications that determine adequate courses of action [68; 91] Examples that computed risk ethics: [18; 23; 29; 30; 43; 68; 69; 70; 82; 94; 96] 	<ul style="list-style-type: none"> Apprehending opinions/ acceptance necessary for effective implementation [3; 14; 16; 19; 23; 24; 25; 50; 63; 78; 80] Machine learning approaches allow SDVs to mimic human ethics (drawing on the "wisdom of the crowd") [22; 34; 44; 68; 69; 87; 94; 96] Examples that computed descriptive ethics: [30; 43; 45; 96]
	Correspondence with reality					<ul style="list-style-type: none"> Uncertainty and variety of possible outcomes (e.g., certainty or degree of injury), especially emergencies in the context of driving, are better dealt with within the framework of risk theory [2; 13; 14; 25; 30; 35; 50; 58; 61; 73; 82; 91] SDV's decisions are made in dynamic conditions of uncertainty [23; 29; 33; 35; 58; 72; 73; 83] 	<ul style="list-style-type: none"> Results from empirical studies are meaningful since individuals usually are not guided by ethical theories but rather by normative standards of their culture [22] Mimicking human behavior (e.g., via machine learning approaches) aligns with the industry's undertakings [88]

Essay II: Applying Ethical Theories to the Decision-Making of Self-Driving Vehicles: A Systematic Review and Integration of the Literature

						<ul style="list-style-type: none">Algorithms of SDVs operate on probabilities rather than on calculations of actual damage [14; 41; 43; 52; 58; 86] <p>Multiple existing sources of risks and variations in the environment (e.g., SDV's driving behavior and technology, state of the road, actions of traffic participants) [13; 24; 32; 33; 34; 38; 44; 51; 55; 57; 63; 66; 70; 72; 73; 83; 91]</p>	
--	--	--	--	--	--	--	--

Table 4: Identified advantages of applying ethical theories to the decision-making of SDVs in past literature

3.3.3 Disadvantages of applying particular ethical theories to the decision-making of SDVs

Similarly, past literature has highlighted several disadvantages of applying particular ethical theories to the decision-making of SDVs. A holistic overview of identified disadvantages in past literature is illustrated in Table 5¹⁵. Again, the disadvantages can be clustered according to social, moral/legal, and functional arguments.

Social disadvantages. Social disadvantages include intransparency of rationale & unpredictability of SDV behavior, generation of suboptimal societal outcomes, and incompatibility with societal preferences/agreements. Concerning the *intransparency of rationale & unpredictability of SDV behavior*, past literature mentioned corresponding drawbacks for four of the theories. Implementing deontological ethics into an SDV's functioning may result in unpredictable behavior for other traffic participants in the sense that it contradicts the common sense or driving style of humans. For example, the key duty of such an SDV would be to conduct literal interpretations of its strict rules and not violate any of its programmed constraints (e.g., adherence to the double yellow lane boundary at any time). Regarding virtue ethics, scholars highlight the challenge of providing insights and predictability of the SDV's decision-making process and consequent behavior, especially when virtues are formed through experience. Additionally, it is questionable what exact outcomes would result from implementing virtues such as courage or pride. The exact behavior and outcomes of utilitarian SDVs are also argued to be less transparent and unforeseeable due to their unrestricted optimization of some utility and complicated calculations that humans may not grasp in real time. Again, the adherence to rigid calculations of such SDVs could lead to unexpected actions that are incomprehensible and inconsistent with human intuition. For example, the SDV would select those actions that fail to restrict individual harm to a moderate level. Lastly, focusing on the Trolley Problem may also entail destructive consequences for society's understanding of SDV behavior in the sense that it artificially attaches too much relevance to an issue (i.e., SDV decisions in dilemma situations) that is rare and unrealistic.

Concerning the *generation of suboptimal societal outcomes*, most selected theories are said to be associated with this disadvantage to some degree. An SDV that is programmed in a deontological manner may create suboptimal overall outcomes due to its objective to strictly adhere to its predetermined rules or if conflicts between its rules arise. For example, such an SDV may undertake dangerous behavior, which ultimately results in decreased levels of safety for

¹⁵ The following text passages summarize overall identified disadvantages without stating specific references. All corresponding references for each disadvantage are indicated in Table 5.

traffic participants in order to conform to its law of never crossing the double yellow. Scholars have discovered that integrating particular virtue ethics considerations (i.e., not endangering families) into SDVs may also result in higher fatalities. Drawing purely on a utilitarian calculus for SDVs can decrease the number of fatalities while the quality of harm is disregarded so that overall worse societal outcomes are generated. Furthermore, if traffic participants know that SDVs function by simple harm minimization, counterproductive incentives and consequences may be created. For example, motorcyclists may choose not to wear helmets to be deemed highly susceptible to harm and thus avoid being targeted in case of an accident. Similar arguments can be made for utilizing contractualism as a guiding logic for SDVs. Namely, if it was known that SDVs function to the advantage of the worst-off, traffic participants may take fewer precautions to avoid constituting the most vulnerable party. Moreover, applying contractualism, instantiated in the Maximin principle, will generate suboptimal societal outcomes quantity-wise since an infinite number of severe injuries is preferred to prevent the death of one person. Mere risk ethics approaches within SDVs (i.e., without counterbalancing utility) can also exhibit suboptimal outcomes for mobility or passengers' comfort in that such SDVs would often decide to stop the care immediately when particular risks are present. The focus on trolley-like training regimes would imply that SDVs, in fact, end up more often in dilemma situations and are ill-prepared for non-dilemmatic scenarios.

Regarding *incompatibility with societal preferences/agreements*, pertinent arguments for five theories are found in past literature. First, as deontological ethics suggest, conceptualizing rules as absolute without consideration of the particular situation at hand cannot always be accepted, especially when dismissing the rules involves positive outcomes such as saving lives. Applying utilitarianism is also not fully compatible with societal preferences since, in some past studies, individuals hesitated to choose and adopt SDVs that decide in a utilitarian manner. Furthermore, society seems divided about what constitutes a proper utility function for SDVs. Therefore, even if society generally agrees on accepting a utilitarian calculus as a legitimate decision criterion for SDVs, the details of a universally valid utility function still need to be worked out. Drawing on a Maximin decision criterion for SDVs may also not be compatible with drivers' preferences since this effectively could entail undesired SDV actions, such as those involving self-sacrifice. Furthermore, even if societal inclinations were investigated in empirical studies, it would not be self-evident that programming SDVs' functioning according to these stated indications entails alignment with societal preferences on a universal level. This is because preferences vary depending on many different variables (e.g., respondents' demographics), the context of the study set-up, or the year in which the investigation is conducted, assuming individuals' preferences may change over time with increased experience with SDVs. Therefore,

scholars have highlighted the inability to find universal, non-contradictory societal preferences and, consequently, to reach a consensual agreement on which ethical rules should determine the functioning of SDVs.

Moral/legal disadvantages. Moral/legal disadvantages include violation of human rights, discrimination/violation of equality, as well as disregard of obligations & responsibility. Past literature references three of the selected theories concerning the *violation of human rights*. Consulting a utilitarian calculus would implicate the problem of incommensurability in that any evaluation of human life and offsetting processes contradict human dignity. Especially the rights of initially innocent traffic participants could be at risk, although a bystander's right to life has stronger legal force than a driver's request to be rescued. Applying risk ethics approaches within SDVs is also said to violate human rights, particularly privacy rights, since its execution would demand the collection of much personal data. Relying on descriptive ethics for the programming of SDVs could also be associated with this drawback since implementing some of the preferences indicated in empirical studies would be incompatible with human dignity and individual rights (e.g., preferences to target a particular age group).

Concerning the *discrimination/violation of equality*, for most of the identified ethical theories, the reasons for how they meet this disadvantage are listed. First, applying deontological ethics in the form of set rules is said to exhibit the potential for discrimination. One absolute moral authority that may be biased would decide which SDV action is right or wrong for every roadway situation. Utilitarian SDVs are reported to violate equality of outcomes in welfare on an individual level since outcomes are maximized for one party, resulting in a reprehension for others. Similarly, such an SDV's logic implicitly prioritizes the majority over minorities, such as one single worst-off traffic participant. Furthermore, a utilitarian calculus that aims to maximize overall safety would imply deliberate and systematic discrimination in that those SDVs, for example, target safer road participants and, thus, penalize them for their safety precautions. Therefore, past literature states that utilitarian SDVs disregard equality and fairness as not everyone's well-being is considered equally. Similarly, to execute the Maximin principle, SDVs will target safer road participants because they have a higher probability of survival. Therefore, SDVs that decide, according to Maximin, also entail discrimination in that they give undue weight to the moral claims of the worst-off. SDVs that function according to risk ethics strategies may distribute risk disproportionately when aiming for improved overall safety. Namely, because vehicle mass highly correlates with crash severity, more crashes with smaller, cheaper cars could occur. This would imply that risk would be implicitly transferred from one party to another without anyone's consent, and relative safety would be up for sale on the market, granted to those buying bigger, more expensive cars. Moreover, implementing some of the preferences stated in

empirical studies or derived from actual driver behavior into the functioning of SDVs could entail the incorporation of unethical preferences, such as discriminating against traffic participants based on unique features.

Concerning the *disregard of obligations & responsibility*, only for three of the considered theories pertinent disadvantages are stated in past literature. Utilitarian SDVs disregard the question of fault and accountability, as they do not make a distinction within their calculus between those individuals who generate road traffic risks and uninvolved parties. Similarly, the Maximin principle may fail to consider legality if assumed that the worst-off traffic participants have made themselves liable to harm as they obeyed traffic laws. Lastly, the traditional Trolley problem disregards who is morally and legally responsible for road accidents, although these facts are essential to constitute the moral permissibility of SDV behavior. Furthermore, the traditional Trolley problem ignores special obligations, such as an automobile manufacturer's duty to protect its customers.

Functional disadvantages. Functional disadvantages include narrow applicability/non-universality, incompatibility & ineffectiveness of implementation, and discrepancy with reality. Concerning the *narrow applicability/non-universality*, five of the selected theories offer corresponding drawbacks to some extent. Deontological ethics is said to offer only limited applicability due to the incompleteness of rules: any set of rules will not encompass all possible situations in advance. Furthermore, situations may arise in which such an SDV cannot adhere to all constraints or rules simultaneously. Therefore, a purely deontological functioning may not be universally practicable and may be limited in flexibly responding to different traffic scenarios. Similarly, scenarios may arise in which implemented virtues such as care and civility contradict each other, for example, when an SDV needs to decide whether to adhere to traffic rules to avoid hitting a pedestrian. Utilitarian logic is not conclusive about concrete SDV action in every traffic scenario. For example, when the number of persons harmed is equal for all possible SDV action alternatives. Similarly, applying the Maximin principle can entail situations in which the SDV cannot make a decision, for example, if the outcomes for two traffic participants are perfectly symmetrical. The Trolley problem does not offer wide applicability in that it only allows the exploration of a narrow solution space (i.e., two possible decision options) and neglects the development of additional solutions. Furthermore, the generated solutions only correspond to specific single cases, denying the generalizability beyond the contemplated case.

Concerning the *incompatibility & ineffectiveness of implementation*, all identified ethical theories are associated with this disadvantage to some degree. Deontological rules are too abstract to be directly adopted within the functioning of SDVs. This is because, in the first place, it is difficult to articulate complex human ethics into a set of rules. Existing laws are not specific

enough to produce concrete instructions for SDV action. Furthermore, depending on an SDV's duty, precise information that is inaccessible is needed, which stands in the way of easy implementation. For example, if an SDV's priority is to avoid injuring humans, full information about the level of harm to humans for all action alternatives is required ex-ante. Regarding virtue ethics, scholars have highlighted the confusion over its exact implementation since no corresponding process for how to make virtuous decisions is stated. Implementing a utilitarian calculus is also argued to be ineffective, as SDVs would need to conduct complex calculations due to the availability of a range of action alternatives and outcomes. Conducting these calculations becomes especially difficult since not everything can be measured in numbers (e.g., the value of human life). Here, also the matter of 'incomparability' plays a role that emphasizes the difficulty of comparing death with major or minor injuries. Additionally, a smooth implementation would require full and precise information to calculate the outcomes of different SDV actions accurately. However, available knowledge of the SDV is inaccessible or incomplete so that, for example, the levels of harm for each road participant or any negative externalities cannot be predicted with certainty. Likewise, full information and precise knowledge of the driving environment and potential outcomes are needed to execute the Maximin principle. Furthermore, contractualism – instantiated in the form of the Maximin principle – is stated to be incompatible with an actual practical implementation due to its difficult and long-lasting calculation. Namely, such an SDV would need to calculate the survival probabilities of each affected party within a few seconds to recognize the worst-off. Like the utilitarian calculus, risk ethics approaches face the issue of determining the value of life, i.e., estimated harm. Moreover, Trolley-preferences methods are not well suited for designing SDV collision algorithms since they are unsolvable, so no ethical implication or solution can be implemented in practice. Furthermore, scholars argue that even if real-life 'Trolley Problem' situations occur, SDVs will not be able to exercise control. This is because, most likely, a dilemma situation arises in the first place due to a total system failure of the SDV, which implicates the inability to adhere to the SDV's initial programming. Therefore, in practice, the emergence of Trolley situations excludes the possibility of an SDV performing trolley-preference methods. Moreover, programming SDVs according to the results of empirical studies is technically difficult since some of the stated preferences and mentioned differentiation factors are hardly detectable (e.g., personal characteristics such as age and gender).

Concerning the *discrepancy with reality*, only for two ethical theories, corresponding arguments are made in that past literature. Rule-based approaches that stem from deontological ethics exhibit a discrepancy with reality as they ignore context-specific information such as the probability of occurrence of future outcomes. Furthermore, scholars have extensively criticized

the Trolley Problem for its simplified and inadequate representation of the real world. This is because the traditional Trolley Problem focuses on implausible, extreme, and rare situations that disregard contextual factors such as potential material damages, interactions between traffic participants, and the existence of uncertainty. In addition, thought experiments presuppose a ‘top-down’ approach to SDV decision-making, in which the decision authority is an unaffected third-person or bystander who prospectively decides about others. The indicated preferences of individuals in such experimental settings thus rather represent superficial claims. In reality, multiple stakeholders, such as citizens, lawyers, and car manufacturers, would negotiate an agreed-upon decision. Lastly, dilemma situations do not correspond to reality as they can be anticipated and are avoidable through proactive prevention.

Essay II: Applying Ethical Theories to the Decision-Making of Self-Driving Vehicles: A Systematic Review and Integration of the Literature

Evaluation criteria of disadvantages		Guiding ethical theories					
		<i>Deontology</i>	<i>Virtue Ethics</i>	<i>Consequentialism/ utilitarianism</i>	<i>Contractualism</i>	<i>Risk ethics</i>	<i>Descriptive Ethics (Trolley Problem)</i>
<i>social</i>	Intransparency of rationale & unpredictability of SDV behavior	<ul style="list-style-type: none"> Strict constraints remove flexibility of SDVs' behavior that may lead to unexpected/extreme behavior contradicting common sense [24; 25; 33; 35; 77; 92] 	<ul style="list-style-type: none"> If virtues are formed through experience, the underlying explanation of SDVs' actions is not fully traceable [29] Unclear how virtues such as courage affect the outcome of a crash or an SDV's functioning [53] 	<ul style="list-style-type: none"> Pursuing optimization may create unforeseeable actions that are less transparent before inspecting its underlying logic; difficult to predict its movement in real time [29; 42] Following rigid calculations lacks common sense and induces actions that contradict human intuition [34; 40; 86] 			<ul style="list-style-type: none"> Emphasis on the Trolley Problem misdirects/overestimates the importance of a negligible issue [25]
	Generation of suboptimal societal outcomes	<ul style="list-style-type: none"> When rules are encoded in a strict deontological manner/ as hard constraints, suboptimal outcomes may be created, e.g., dangerous driving behavior to adhere to rules, increased number of fatalities, high costs due to property damage [18; 25; 29; 33; 34; 36; 37; 57; 79; 100] If rules conflict, SDVs may be unable to produce action decisions and execute suboptimal fallback plans (e.g., brake) [36; 40; 53; 65; 74] 	<ul style="list-style-type: none"> With the enactment of particular virtues, virtue ethics approaches may result in a higher number of fatalities [53] 	<ul style="list-style-type: none"> Counterproductive incentives/consequences are created if it is known that SDVs function according to overall harm minimization and, thus, may entail disadvantages for those taking safety precautions, e.g., motorcyclists may choose not to wear helmets to avoid being targeted, manufacturers may face litigation from SDV owners [2; 5; 19; 22; 23; 33; 40; 53; 57; 58; 67; 78; 86; 93; 95] Focus on quantity instead of quality/nature of harm [86] 	<ul style="list-style-type: none"> Quantity of harm disregarded; Maximin principle prefers an infinite number of severe injuries to prevent a single person's death [25; 29; 44; 48; 54] Counterproductive incentives/consequences are created if known that SDVs function according to the Maximin principle/in the advantage of the worst-off, e.g., individuals would stop wearing helmets and buying safer cars [48; 54] 	<ul style="list-style-type: none"> Mere minimization of risk without any utility considerations may result in counterproductive actions that inhibit functionality, e.g., stopping the SDV immediately may have negative effects on mobility, passenger comfort, and energy consumption [91] 	<ul style="list-style-type: none"> Self-fulfilling prophecy: If trolley-like training regimes are adopted, SDVs would end up more often in such scenarios or do poorly in non-dilemmatic scenarios [4]
	Incompatibility with societal preferences/agreements	<ul style="list-style-type: none"> The absolute nature of rules and disregard of circumstances cannot always be accepted, especially if it involves saving lives [22; 47] and when the ratio between good and bad outcomes increases [84] 	<ul style="list-style-type: none"> No agreement on universal moral virtues that should guide SDV's reasoning process [81] 	<ul style="list-style-type: none"> No agreement on what constitutes an acceptable utility function/a universally valid utilitarian calculus [11; 25; 66] Utilitarian calculus does not correspond to human intuitive decisions [85] Missing true willingness to adopt a utilitarian calculus (as it may imply self-sacrifice) [2; 5; 6; 9; 10; 22; 25; 27; 29; 33; 34; 53; 56; 57; 59; 64; 67; 75; 78; 80; 88; 95; 96; 97; 99; 100; 101] 	<ul style="list-style-type: none"> Maximin is not always (rationally) acceptable to individuals, e.g., if the outcomes involve self-sacrifice or disregard probabilities [12; 44; 48] 		<ul style="list-style-type: none"> Indicated preferences for SDVs' particular functioning contradict each other; limited consistency in decision-making [2; 6; 9; 17; 19; 25; 29; 36; 56; 60; 61; 70; 71; 82; 84; 101] Preferences change over time, e.g., with SDV experience [12; 22; 41; 44; 72] Stated preferences depend on various variables and the study context, e.g., the type of victim, probability/ risk considerations, severity of outcome, respondents' demographics [1; 2; 3; 5; 6; 9; 14; 16; 22; 25; 27; 28; 29; 31; 47; 54; 56; 60; 61; 63; 64; 70; 75; 78; 80; 81; 84; 89; 90; 97; 101] Inability to reach a consensual agreement on what moral principles should be shared / what general rule should apply for SDVs [1; 5; 19; 20; 22; 23; 24; 49; 50; 58; 60; 67; 70; 81; 97; 100]
<i>moral/ legal</i>	Violation of human rights			<ul style="list-style-type: none"> Pure utilitarian calculus neglects that killing is worse than letting die, e.g., an innocent bystander's right to life has stronger legal force than someone's request to be rescued [26; 42; 49; 54; 57; 78; 95] 		<ul style="list-style-type: none"> Potentially entails privacy concerns due to the necessity to collect much data; massive and systematic breach of civilian privacy rights [13] 	<ul style="list-style-type: none"> Implementing some of the stated preferences from past studies would be incompatible with human dignity and individual rights [6; 15; 21; 90]

Essay II: Applying Ethical Theories to the Decision-Making of Self-Driving Vehicles: A Systematic Review and Integration of the Literature

<i>functional</i>				<ul style="list-style-type: none"> Problem of incommensurability; human dignity contradicts any evaluation of or setting off human life [17; 21; 24; 29; 31; 32; 35; 40; 42; 44; 51; 57; 62; 66; 71; 80; 86] 			
	Discrimination/ violation of equality	<ul style="list-style-type: none"> One absolute moral authority/rule that may be biased would decide what is right and wrong in every roadway situation [17; 34; 47] 		<ul style="list-style-type: none"> Majority is prioritized over minorities in road traffic [15; 79] Deliberate and systematic discrimination/sacrifice based on personal features, e.g., SDVs will target safer road participants/penalize them for safety precautions [2; 6; 11; 25; 31; 35; 40; 44; 53; 57; 58; 66; 77; 81; 86; 91] Welfare of individuals is considered unequally [6; 20; 29; 31; 83] Outcome maximization for one results in loss for others [20; 25; 30; 80] 	<ul style="list-style-type: none"> Unfair/unjustified discrimination and sacrifice, e.g., to execute Maximin principles, SDVs will target safer road participants on the grounds that they have a higher probability of survival (here survival probabilities act as a proxy for harm) [29; 54] Maximin principle gives undue weight to the moral claims of the worst-off [29; 48] 	<ul style="list-style-type: none"> Implicitly uneven risk distribution when aiming for overall safety, e.g., as vehicle mass highly correlates with crash severity, more crashes with smaller/cheaper cars could occur; relative safety is up for sale, granted to those who buy big/costly cars [7; 36; 37] Risk transfer from one party to another without anyone's consent [36] 	<ul style="list-style-type: none"> Implementing some of the stated preferences in past studies would entail discrimination by sacrificing based on personal characteristics [3; 6; 15; 21; 31; 62; 81; 90] Implementing some of the stated preferences in past studies or derived from driver behavior would entail the incorporation of unreflected, intuitive, and unethical preferences; observing individuals will not teach an SDV what is ethical, but what is common [19; 22; 23; 25; 29; 31; 34; 65; 68; 70; 81; 82; 83; 88; 94; 100]
	Disregard of obligations & responsibility			<ul style="list-style-type: none"> Responsibility/the question of fault is disregarded, i.e., no difference between those that generate road traffic risks and uninvolved parties [25; 26; 31; 42; 86] 	<ul style="list-style-type: none"> Maximin fails to take account of legality, e.g., when drivers have made themselves liable for some harm by intentionally breaking the law [44] 		<ul style="list-style-type: none"> Trolley problem disregards obligations, e.g., a manufacturer's duty to protect customers [50; 60; 91] Trolley problem disregards moral and legal responsibility, e.g., for a road accident [4; 25; 29; 42; 50; 60; 61; 66; 72; 73]
	Narrow applicability/ non-universality	<ul style="list-style-type: none"> Incompleteness of rules; any set of rules does not encompass all possible situations [2; 25; 29; 32; 35; 40; 56; 79; 80; 81; 94; 100] Poor flexibility [25] Scenarios may arise where SDVs cannot fulfill all constraints/rules at once [2; 19; 25; 32; 33; 40; 74; 76; 88; 92; 94; 100] 	<ul style="list-style-type: none"> Situations may arise in which implemented virtues conflict with each other [32] 	<ul style="list-style-type: none"> Does not provide an answer to what to do when the number of persons harmed is equal for all possible outcomes [65] 	<ul style="list-style-type: none"> Maximin may entail situations in which it is unable to make a decision, e.g., if the outcomes for two traffic participants are perfectly symmetrical [29; 54] 		<ul style="list-style-type: none"> Trolley-preferences method focuses on narrow solution space; it has no mechanism for discovering new options available to SDVs [38; 42; 82; 98] Trolley problem considers one single-cases, thus, neglects the issue of aggregation [22; 25; 41; 58; 83]
	Incompatibility & ineffectiveness of implementation	<ul style="list-style-type: none"> Rules that are available to be adopted are too abstract to be directly adopted; difficulty in articulating complex human ethics into a set of rules [29; 33; 35; 40; 94; 100] Existing laws are not specific enough to guide actions in SDVs [34] Full information/precise information needed, e.g., what actions result in harm if an SDV's duty is to avoid injuring humans [33; 34; 64] 	<ul style="list-style-type: none"> Unclear implementation as no specific process for making virtuous decisions or how to implement virtues is offered [32; 40] 	<ul style="list-style-type: none"> Complex calculation due to the availability of a range of potential outcomes [25; 34; 78] Not everything can be measured in terms of numbers or a linear manner, e.g., it is difficult to establish a standardized scale for the value of human life [5; 21; 25; 47; 86; 95] Full information is needed to accurately calculate outcomes; but inaccessible, uncertain, and incomplete, e.g., prediction of harm for all road users and negative externalities [5; 11; 17; 21; 23; 25; 32; 33; 34; 44; 51; 64; 78; 81; 86; 88; 91; 92; 94] 	<ul style="list-style-type: none"> Difficult and long-lasting calculation, e.g., SDVs would need to calculate the survival probabilities of each affected party within a few seconds [48; 51; 54] Full information/precise knowledge is needed but inaccessible due to driving environment with unforeseen and uncountable variables [17; 25; 51; 77] 	<ul style="list-style-type: none"> Calculating the value of life is difficult, e.g., how to weigh up a serious injury that entails lifelong disabilities with death [29; 36] Full information/precise knowledge is needed but inaccessible [13; 29; 51; 52; 73] 	<ul style="list-style-type: none"> SDVs will not be able to exercise control/make decisions in real-life 'Trolley Problem' situations, either in time or at all, due to a total system failure that led to the situation arising in the first place [25; 39; 41; 83] Trolley problem is unsolvable, and no ethical implication/solution is derived [4; 6; 38; 42; 71; 81; 98] Discrimination based on personal characteristics is difficult since some differentiation factors are hardly detectable [13; 21; 51; 57] Moral understanding cannot be modeled computationally [19]
	Discrepancy with reality	<ul style="list-style-type: none"> Rule-based approaches ignore context-specific information such as 					<ul style="list-style-type: none"> Trolley problem focuses on implausible, extreme, and rare situations [2; 12; 22; 25;

Essay II: Applying Ethical Theories to the Decision-Making of Self-Driving Vehicles: A Systematic Review and Integration of the Literature

		the probability of occurrence of future outcomes [29; 77]					<div>36; 38; 39; 50; 52; 54; 60; 61; 70; 71; 80; 82; 83; 88; 100]</div> <ul style="list-style-type: none">• Trolley problem assumes 'top-down' decisions of an unaffected third person about others [41; 42; 50; 73]• In reality, dilemma scenarios can be anticipated and proactively avoided [13; 38; 42; 46; 58; 66]• Trolley problem is an idealized representation; neglect of contextual matters, e.g., material damages, interactions between participants; different time perspectives [4; 10; 13; 24; 25; 36; 41; 42; 53; 54; 58; 64; 66; 72; 73; 78; 80; 81; 85; 91; 98]• Trolley problem assumes a limited number of certain outcomes, risk/probability is disregarded [2; 4; 13; 14; 19; 22; 25; 29; 35; 36; 41; 42; 50; 52; 58; 61; 64; 65; 70; 72; 73; 82; 83; 88; 91; 98]• Indicated preferences provide limited insight into what society actually deems acceptable, e.g., due to social desirability bias [5; 17; 24; 36; 39; 41; 50; 59; 63; 64; 65; 71; 78; 81; 83; 85]
--	--	---	--	--	--	--	--

Table 5: Identified disadvantages of applying particular ethical theories to the decision-making of SDVs in past literature

3.3.4 Combined theories, additional considerations/principles & elaborate approaches for the decision-making of SDVs

In the past, scholars acknowledged the insufficiency of relying on any single ethical theory when programming SDVs and consequently pressed the need to create a ‘mixed’ algorithm that combines elements from various approaches (e.g., Brändle & Schmidt, 2021; Goodall, 2014b; Hübner & White, 2018; Thornton et al., 2017; Wang et al., 2020). This way, an SDV’s functioning would not be “tightly bound to traditional moral theory, or any single conception of the good” (Evans, 2020; p.3295). Such combined approaches could account for various ethical concerns, capture the flexibility of human moral judgment, and consequently, are more likely to achieve widespread societal acceptance (Dubljević, 2020; Hübner & White, 2018). After all, it seems individuals’ preferences for an SDV’s decision-making are more nuanced instead of conforming to one single ethical theory (Ryazanov et al., 2023). Similarly, policymakers from, for example, the European Commission have called for the combination of several ethical principles (Geisslinger et al., 2023). Correspondingly, researchers have proposed hybrid combinations of particular theories and additional considerations/principles as well as sketched the first elaborate ethical decision processes for SDVs, as illustrated in Table 6¹⁶.

Hybrid combinations of theories that have been proposed are, for example, mixing various traditional normative theories among themselves or mixing them with descriptive ethics or risk ethics. Concerning the mix of normative theories, scholars have suggested combining utilitarianism (e.g., in the form of harm minimization) with deontological ethics (e.g., in the form of side constraints), adopting a deontic logic of utility maximization, implementing threshold deontology to account for ‘bad enough’ outcomes or proposed a general social welfare function that integrates aspects of utilitarianism, contractualism and efficiency considerations. Concerning the combination of normative theories with descriptive ethics, researchers propose approaches in which society’s preferences generated through empirical studies, inclusive deliberation, or machine learning are included in SDVs’ decision functions (e.g., in the form of moral weights for particular considerations or restrictions on what outcomes are to be taken into account). Furthermore, scholars have proposed the adoption of risk as a critical value to be considered when contemplating normative theories by, for example, utilizing risk as a factor to be optimized or counterbalanced in (utilitarian) cost functions.

Additional considerations and principles that have been highlighted in past literature include the admissibility of non-arbitrary discrimination, the notion of reasonableness, constructing equal chances for affected parties, randomizing decision actions (especially when ethical theories are

¹⁶ Table 6 indicates the corresponding references for each suggested solution.

inconclusive), granting priorities to particular parties (e.g., the owner of the SDV or individuals outside the vehicle), as well as distribution strategies that are sensitive to a parties' legal compliance and responsibility for a causing dangerous situation. Furthermore, scholars indicate that legitimate moral action and distribution decisions for SDVs may depend on the situation at hand (i.e., hazard situation vs. non-hazard case) and may benefit by drawing inspiration from existing distribution strategies in other fields. For example, touching on utilitarian considerations, the area of healthcare offers potential value-of-life estimates, or the literature on radiation exposure introduces the idea of thresholds/individual dose limits.

Lastly, past literature has also put forward more *elaborate approaches*¹⁷ that suggest continuous decision-making processes for SDVs when engaging in trajectory selection. For one, the Data theories method – based on the situation at hand – determines all action options for an SDV and the probability of their consequences. Afterward, the SDV will “either choose an option that is favored by at least one plausible ethical theory or will choose an option that is favored by the best theory of how to compromise between different ethical theories” (Robinson et al., 2021b; p.2). Furthermore, the Ethical valence theory “resembles a pluralist form of act consequentialism which abides by contractualist constraints and principles” (Evans et al., 2020; p.3291). Namely, the theory roughly consists of four determinants: (1) high-level duties (e.g., “The lives of the passenger(s) must not be put in harm’s way”) that need to be fulfilled or otherwise initiate the subsequent ethical deliberation process, (2) ethical valences of (the claim of) each road user that is determined by a socially acceptable classification or hierarchy (e.g., pedestrians have higher valence than vehicle passengers), (3) expected harm serving as a threshold and (4) moral profiles (e.g., risk-averse altruist vs. threshold egoist) that dictate to what extent SDVs are most sensitive to which claims (Evans et al., 2020). Moreover, the Expected moral value approach juxtaposes two actions, A (“crash into roadster”) and B (“not crash into roadster”), and weighs the moral value of each action with one’s credence (i.e., subjective probability) in the rightness of conducting a particular action (Bhargava & Kim, 2017).

In addition to these elaborate theoretical decision-making processes, scholars have put forward explicit ethical trajectory-planning algorithms. For example, Németh (2022) proposed a route selection method that consists of a quantitative and a qualitative evaluation layer. In line with consequentialist viewpoints, the SDV would first engage in a quantitative calculation to find the route with the lowest cumulative probability for a critical conflict. If the remaining routes are

¹⁷ Note: The approaches that are summarized here are not equivalent to the number of publications that showcase concrete technical implementation (compare Figure 22). For example, while D’Souza et al. (2022) or Jiang et al. (2023) simulate and compare how different ethical theories or principles can be computed, they do not conclude their articles with the proposition of one comprehensive decision-making process. Therefore, articles like these are not listed here.

still expected to result in serious or fatal injuries, additional ethical principles (such as sparing uninvolved traffic participants and enacting random choice) are consulted in the qualitative evaluation layer (Németh, 2022). A similar control design is proposed in Németh (2023), in which, however, stopping the vehicle (instead of choosing a trajectory randomly) can be conducted as a last resort. Also, Thornton et al. (2016) introduced a decision process in which the control algorithm builds an optimization function for each feasible path. This optimization function incorporates objectives such as path tracking or vehicle occupant comfort as priorities and corresponds to additional constraints such as obstacle avoidance. Based on these calculations, the optimal path is then determined and chosen (Thornton et al., 2016). Geisslinger et al. (2023) developed an ethical trajectory-planning algorithm that samples potential trajectories and the corresponding risks they pose to all traffic participants. Sampled trajectories are then categorized into four validity levels, in which trajectories are declared ‘invalid’ that, for example, surpass a maximum acceptable risk. For trajectories in the highest validity level (i.e., ‘valid’), an ethical cost function is calculated that consults the minimization of the overall risk, the priority for the worst-off, equal treatment of people, and responsibility considerations. Eventually, the trajectory is selected for execution with the highest validity level and the lowest calculated cost (Geisslinger et al., 2023). Other scholars have suggested an algorithm that allows personal ethics settings (Wang et al., 2022b). Namely, potential trajectories within the road boundary are sampled. These are then evaluated and selected based on their consistency with public ethical preferences (e.g., concerning the relative importance of injury levels and number of casualties of road users), which are derived from user studies on dilemma situations (Wang et al., 2022b). In 2020, Wang et al. proposed a Lexicographic Optimization based Model Predictive Controller for ethical decision-making, which collects obstacle and environment information as a first step. After having determined the possible action field and any road boundaries, the crash severity for all existing obstacles (e.g., pedestrian, cyclist, vehicle) is calculated upon which the obstacle’s priority is determined. The SDV then decides on what is the best action based on the information from the potential field, the obstacle priority, and other constraints or objectives such as regulation and trajectory following terms or the avoidance of obstacles to the greatest extent (Wang et al., 2020). Islam and Rashid (2018) suggested “[a] priority queue based min-heap sorting algorithm” (p.1), in which, at the time of an imminent collision, the survival probabilities for all traffic participants are calculated for each incident (i.e., trajectory). Generally, the incident with the least chance of survival should be avoided. To make a final action decision, the SDV additionally generates a score for each incident by categorizing the lives of affected traffic participants according to some given parameters such as age and gender. The SDV then selects and executes the incident with the combination of the lowest score and survival probability (Islam & Rashid, 2018).

Similar to the previously stated elaborate theoretical approaches, these identified ethical trajectory-planning algorithms, on a general level, seem to pursue the following sequence of steps: (1) Analysis of the traffic situation at hand and sampling of all possible trajectories, (2) Evaluation of possible trajectories based on balancing or prioritizing certain objectives and constraints, (3) (Sequential) trajectory elimination until action decision and final trajectory is determined.

Hybrid combinations	<i>Mix of traditional normative theories</i>	<ul style="list-style-type: none"> • Moral rules as constraints and utilitarianism in the form of weighing costs and decision options [33; 91; 92] to impel harm minimization [25; 42; 70] • Deontic logic of utility maximization [87] • Threshold deontology in that there exists a threshold at which the magnitude of harms (i.e., consequentialist considerations) override deontological constraints [84] • General social welfare function that integrates aspects of utilitarianism, contractualism, and efficiency considerations [20]
	<i>Traditional normative theories & descriptive ethics</i>	<ul style="list-style-type: none"> • Agent-deed-consequence model that (based on empirically investigated moral weights) counterbalances, e.g., a negative deed (e.g., breaking the law) with an agent's good intention and beneficial consequences [19] • Utilitarian logic that aims to minimize overall injury severity, while the range of collision outcomes that can be considered in the first place correspond to participants' indications in surveys [75] • Inclusive, public deliberation that allows the development of an ethical goal function that reflects the norms and values of the broader public [79; 81]; and buyers' or manufacturers' ethical intuitions (such as the considerations of a victim's age) [40; 43] • Adoption of machine learning methods to formulate ethical rules based on human behavior in real-world/simulated crash scenarios, while normative theories act as behavioral boundaries [44]
	<i>Traditional normative theories & risk ethics</i>	<ul style="list-style-type: none"> • Combination of Bayesian, equality, and the Maximin principle – each with a particular weighting factor – with risk as the key distribution value [29] with the additional constrain of a maximum acceptable risk threshold [30] • Counterbalancing risk and utilities such as mobility to counteract the complete stop of SDVs that is the safest option in most cases [91] • Utilitarianism with the utility of risk to be minimized for everyone [7], complemented by additional constraints or priority lists [94]
Additional considerations /principles	<i>Non-arbitrary discrimination</i>	<ul style="list-style-type: none"> • Non-arbitrary discrimination is justified as it is made due to a morally relevant reason (e.g., airlines discriminating against blind pilots) [31]
	<i>Notion of reasonableness</i>	<ul style="list-style-type: none"> • For example, accepting minimal chances of harming an individual (e.g., 1%) if it is expected to result in saving another person with certainty [91]
	<i>Greatest equal chances/strict equality</i>	<ul style="list-style-type: none"> • Construction of a weighted lottery between alternatives, where the weightings are fixed to ensure that the affected parties receive the greatest equal survival probabilities [7; 15; 30; 45; 48] or equal statistics-based road risks [86]
	<i>Randomization</i>	<ul style="list-style-type: none"> • Random selection of which side to be sacrificed in a dilemma situation [17; 25; 40; 66; 78; 100; 101] when underlying theory (e.g., Maximin principle) is inconclusive [54] or when traffic participants hold equal degrees of responsibility for a particular traffic situation [21]
	<i>Prioritization</i>	<ul style="list-style-type: none"> • Protection of SDVs' owners ('passenger first') [5; 7; 21; 28; 88; 94; 101]; 'ethical egoism' [40; 45; 53] • Prioritizing pedestrians over SDV passengers; 'altruistic mode' [21; 45; 69]
	<i>Sensitivity to law & responsibility</i>	<ul style="list-style-type: none"> • Minimization of total harm, where the level of allocated harm for all individuals is sensitive to their compliance with the law [30; 44] • Consideration of the amount of a party's responsibility for a dangerous situation [91; 95]; e.g., drivers may be obligated to absorb most of the harm since they introduce the risk to the road traffic in the first place [21; 30; 57; 63]

Essay II: Applying Ethical Theories to the Decision-Making of Self-Driving Vehicles: A Systematic Review and Integration of the Literature

	<i>Situation-adjusted distribution</i>	<ul style="list-style-type: none"> Principles for distribution depend on the situation at hand (i.e., hazard vs. non-hazard situation) [15; 16; 80; 81]; certain thresholds or obligations are imposed only under certain conditions [30; 68; 69; 79; 87]
	<i>Distribution strategies from other fields</i>	<ul style="list-style-type: none"> Existing standards in healthcare resources allocation/organ donation, such as the number of life-years saved as potential value-of-life estimates [11; 38] Compliance with individual dose limits (thresholds) of radiation exposure should take precedence over total radiation exposure [37]
Elaborate decision processes	<i>Data theories method</i>	(1) Composition of all SDV's potential ethical choice scenarios (2) Determination of all action options and probability of their consequences (3) Based on ethical theories, determination of obligatory, permissible, and prohibited actions (4) Programming SDV to choose the option that corresponds to most/best ethical theories [82]
	<i>Ethical valence theory</i>	(1) Typification of dilemma situation: if SDV cannot fulfill certain duties (2) Determination of strength of ethical valences (i.e., hierarchization of road users based on their claims) (3) Calculation of expected harm for action alternatives by determining the difference of velocity between road users (4) Final action selection dependent on self-determined moral profiles [23]
	<i>Expected moral value approach</i>	(1) Considering two morally analogous actions, A and B (2) Determination of one's credence in the two propositions A and B (3) Identification of the relative difference in the magnitude of the moral value between the two propositions (4) Calculation & selection of the difference between the expected values of both propositions (e.g., credence x moral value) [8]
	<i>Ethical trajectory-planning algorithms</i>	(1) Calculating the total probability of critical conflicts for each route (2) Finding the one route with minimal aggregated probability for a critical conflict (3) If remaining routes result in serious or fatal injury, routes are to be selected that do not involve 'new' participants (4) If remaining routes still result in serious or fatal injury, the final route is selected by random choice [69]
		(1) Objectives (such as safety, comfort, and energy) are formed and balanced in a cost function (2) Constraints on the reachable trajectories are added (i.e., handling all traffic participants equivalently, prohibition to move on the sidewalk) (3) If it is not possible to find a feasible trajectory, the SDV must stop [70]
		(1) Feasible trajectories are determined (2) Optimization function is calculated for each trajectory by incorporating objectives/priorities (e.g., path tracking, passenger comfort) as well as constraints (e.g., obstacle avoidance) (3) Option with lowest cost resulting from the optimization function is selected [92]
		(1) Potential trajectories are sampled, and risk values for every road user in each trajectory are calculated (2) Sampled trajectories are categorized into validity levels (e.g., based on factors such as their breach of a risk threshold) (3) For trajectories in the highest validity level, an ethical cost function is calculated (4) Selection and execution of final trajectory from the highest validity level and the lowest calculated cost [30]
		(1) Personal ethics settings are modeled using the inclination of evasive steering (IES) with injury level and number of casualties of road users as sensitive factors (2) Candidate trajectories are sampled (3) The candidate trajectory with the largest IES value is chosen [96]
		(1) Data is collected from the traffic situation at hand to generate the potential field, and existing road boundaries (2) Risk is analyzed through potential crash severity (3) Obstacle's priority is determined by evaluating the potential crash severity (4) Priorities are built over multi-constraints and objectives (5) Decision-making outputs and control commands to the SDV are generated [94]
		(1) The probability of survival for every incident is calculated (2) A score for each incident is generated based on values such as age and gender (3) An incident with a combination of the lowest score and survival probability will be selected [43]

Table 6: Overview of suggested combined theories, additional considerations/principles, and elaborate decision processes in past literature

3.4 Discussion

After having illustrated a holistic literature overview of the applicability of various ethical theories to the decision logic of an SDV, this section will summarize key takeaways and implications (3.4.1) and point out critical remarks of this review and an updated research agenda (3.4.2).

3.4.1 Key takeaways and implications from past literature

Theoretical articles dominate research, but empirical studies and concrete implementation suggestions are catching up. As can be deduced from the structural analysis, articles that are conceptual and do not refer to the technical implementation of the discussed theories make up the majority of publications in past literature on autonomous driving ethics (see Figure 21 and Figure 22 in Appendix B of Essay II – Descriptive/structural analysis of the literature). However, articles that propose and illustrate concrete suggestions for how to technically implement ethical theories into SDVs have been increasing over the years (see Figure 22 in Appendix B of Essay II – Descriptive/structural analysis of the literature). Similarly, empirical studies such as quantitative surveys, experiments, and simulations are more prevalent nowadays than the efforts at the beginning of the respective research field (see Figure 21 in Appendix B of Essay II – Descriptive/structural analysis of the literature). After all, incorporating ethical considerations into software and conducting tests in simulations will shed additional light on their technical feasibility and resulting consequences (e.g., in terms of occurring accidents) (Geisslinger et al., 2023). More recently, scholars have started to introduce and test explicit ethical trajectory-planning algorithms (e.g., Németh, 2023; Wang et al., 2022b) that thereby demonstrate the technical feasibility of a pluralistic consideration of normatively relevant factors and theories. More generally, an upward trend toward researchers' focus on developing more practical, integrated, and sophisticated approaches in the field of autonomous driving ethics can be recorded (see Figure 23 in Appendix B of Essay II – Descriptive/structural analysis of the literature). This development corresponds to manufacturers' "more pragmatic, technology-infused" approach (Martinho et al., 2021; p.571), potentially allowing the SDV industry to seriously account for ethical considerations in the programming of SDVs' decision-making.

Each ethical theory exhibits social, moral/legal, and functional advantages and disadvantages; thus, hybrid combinations are necessary to counteract tradeoffs. As illustrated by this review, a mix of social, moral/legal, and functional considerations exist when evaluating the applicability of various ethical theories to the decision-making of SDVs. As further demonstrated, none of the identified theories on their own are associated with all benefits while being unaffected by any drawbacks. Therefore, this article does not aim to yield knockout

arguments against or in favor of any single theory, but it strives to help researchers, companies and, policymakers make more informed and balanced decisions when (mandating how to) program(ming) an SDV's decision-making. Namely, the provided matrices on advantages and disadvantages (see Table 4 and Table 5) can serve practitioners as a helpful reference by highlighting essential evaluation criteria for the assessment of applying (combinations of) ethical theories. Furthermore, the provided matrices point towards potential negative outcomes of adopting particular ethical theories for which countermeasures need to be established, as well as towards positive outcomes of adopting ethical theories that may serve as fitting remedies. Based on these insights, practitioners may succeed in building an integrated approach that counterbalances tradeoffs and attenuates the disadvantages of particular underlying theories. To elaborate, any proposed hybrid combination (see Table 6) that integrates the identified theories could be analyzed in part from a social, moral/legal, and functional perspective. Namely, one could draw on the indicated advantages and disadvantages, which relate to the individual ethical theories that this certain combination is composed of. For example, a combination based on, amongst others, risk ethics, utilitarian logic, and the Maximin principle (e.g., Geisslinger et al., 2021; Geisslinger et al., 2023), may still in theory exhibit the disadvantages that relate to utilitarianism (e.g., prioritizing the majority over the minority) and the disadvantages that relate to contractualism (e.g., giving undue weight to the worst-off despite low collision probability). As these ethical theories are integrated, it could be expected they interact with each other in a way that they attenuate each other's disadvantages by serving as a counterweight or countermeasure¹⁸. A summarizing model of integrating the various ethical theories is sketched in the following paragraph, which is to be complemented by future empirical studies that address the concerns and open questions listed in Section 3.4.2.

Each ethical theory can be practically applied to the ethical decision-making of an SDV.

As previously illustrated in Table 3, derived from the implementation possibilities and computed examples indicated in Table 4, as well as the ethical trajectory-planning algorithms in Section 3.3.4, it can be deduced that all identified theories could (to varying degrees) be translated into graspable guidance for either the SDV's ethical decision-making or the process of determining that decision-making. To elaborate, *deontological ethics* can be applied in the form of rules as *soft constraints and hierarchical orders* to somewhat bind SDV behavior. *Virtue ethics* can serve

¹⁸ Although the matrices (Table 4 and Table 5) can be utilized as a first indication of potential advantages and disadvantages of particular integrated decision-making processes, further (empirical) research is needed to investigate to what extent/whether the disadvantages of the individual ethical theories indeed balance each other out when being integrated. Furthermore, it needs to be acknowledged that when ethical theories are integrated, they may depart from their original conception on an individual level. Such new understandings of a previously 'strict' ethical principle, such as Maximin, must be made transparent (Kirchmair & Lando, 2023).

as a basis for establishing *signals in reinforcement learning* or for determining, for example, the strength of applied rules and costs for different types of traffic participants based on their role in fulfilling certain virtues. *Consequentialism/utilitarianism* is applicable in a *cost function* that draws on various utilities to determine the trajectory that maximizes the cost-benefit ratio. *Contractualism* is transferrable in the *Maximin principle* to guide SDV decision-making to minimize the most significant harm, or it can be utilized as a *methodology in empirical studies* to decide on legitimate/acceptable actions, for example, behind the veil of ignorance. *Risk ethics* provides a more concrete *measure of calculating and balancing outcomes* in traffic via the variables of collision probability and estimated harm. It can also serve as a constraint in the form of *risk thresholds*. *Metaethics* can be drawn on in the form of a *high-level perspective or embedded culture* from which fundamental ethical issues are raised and considered that ultimately shape the subsequent SDV programming process and approach. Lastly, *descriptive ethics* can generate *insights into human moral preferences*, which can potentially be incorporated into the SDV's functioning. Figure 18 summarizes and integrates these application possibilities at a macro-level¹⁹. First, a cost function that integrates different principles, utilities, and theories can be built (*layer 1*). On top of this function, guardrails in the form of rules/constraints and thresholds that must not be exceeded can be utilized to adjust and restrict the optimization function and the trajectory alternatives that may be considered in the first place (*layer 2*). To determine the numerical figures and concrete considerations for layers 1 and 2, contractual approaches, additional considerations/principles, descriptive ethics, and virtue ethics can be consulted (*layer 3*).

Overall, the purpose of this model is to show how the different ethical theories can be integrated into the decision-making of SDVs so that they decide based on ethical considerations. This will become relevant when an SDV has to decide between different trajectories. According to our model, the SDV would then evaluate its options based on its prospect of optimizing certain utilities, complying with certain rules, and abiding by certain risk thresholds (see layers 1 and 2). What (the weight or hierarchy of) these utilities, rules, and thresholds concretely are would be informed by additional ethical theories (see layer 3) and metaethical considerations. For example, one rule or utility derived from contractualism could be the prioritization of the most vulnerable traffic participant. Furthermore, the SDV could consider utilities such as equality among the traffic participants or instigate a rule of adjusting its behavior to the situation at hand (e.g., safety is to be prioritized over other utilities such as comfort in hazard situations). Moreover, mimicking

¹⁹ Figure 18 is not to be understood as the chronological order in which an SDV would execute its decision-making process. Instead, it illustrates how the different theories can be applied at all (e.g., as rules/constraints). For example, during actual SDV decision-making, certain rules/constraints could be enforced before the optimization process by eliminating unwanted trajectory alternatives.

individuals' preferences or using certain virtues as positive signals in reinforcement learning may yield further specific rules. All these layers and decision processes are embedded in metaethical considerations that prevail in the field of autonomous driving ethics.

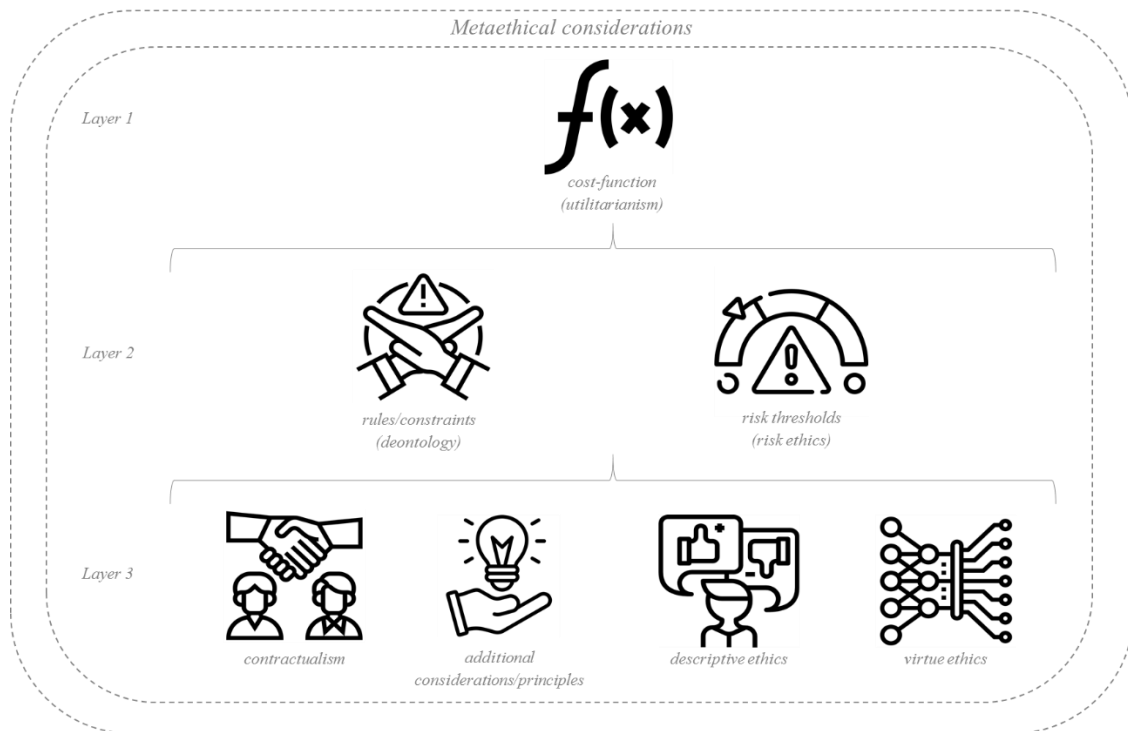


Figure 18: Summarizing model of applying and integrating ethical theories to SDVs' decision-making

Risk ethics represent an integral part of the ethical decision-making of SDVs. If keen to move away from purely theoretical considerations of integrating ethics into the decision-making of an SDV, it appears essential to draw on risk ethics approaches as a critical feature or complement. Namely, as highlighted in past literature, risk ethics aligns most with reality, i.e., with how SDVs operate in traffic (see Table 4). For this reason, industry players such as Google may have considered risk distribution strategies in their patents guiding vehicles' lateral lane positioning in the past (Dolgov & Urmson, 2014). More recently, policymakers have also called for risk management as a guiding imperative and established corresponding frameworks. For example, on a general level, the European Commission (2021) proposes a risk-based approach to govern artificial intelligence. More specific to SDVs, it is suggested to "manage dilemmas [emerging in street traffic] by principles of risk distribution" (European Commission, 2020; p.7). From our research, insights can be deduced about what 'risk-based approaches' more specifically entail in (the process of programming) an SDV's ethical decision-making. To elaborate, risk considerations translate to the optimization function by not counterbalancing certain outcomes/utilities but rather contemplating estimated harm and collision probability. Moreover,

risk considerations can act as constraints in the form of risk thresholds (i.e., certain levels of estimated harm or collision probability that are not to be exceeded). Lastly, risk considerations can be adopted within empirical studies to generate preferences (for underlying principles/theories) for SDV decision behavior that are internally and externally valid, i.e., in road traffic that is characterized by uncertainty. Therefore, risk ethics can play a central role in the development of SDVs' ethical decision-making, i.e., throughout the layers that are illustrated in Figure 18.

3.4.2 Critical remarks and updated research agenda

Some identified advantages and disadvantages are not specific to SDVs and may apply to other technological innovations as well. This research has identified the advantages and disadvantages of applying particular ethical theories to SDVs' decision-making. When considering some arguments for or against a particular theory, it can be assumed that they may similarly apply to other technologies. For example, the insight that rule-based approaches are too strict and ignore context-specific information corresponds to the technical applicability of deontological arguments on a general level. Therefore, some of the described findings may be relevant to and generalizable for various other technologies, offering a first understanding of which theories are implementable into the logic of technology and what corresponding advantages and disadvantages would be. Similarly, the established matrix (i.e., Table 4 and Table 5) could be utilized in future research to analyze (and extend) the applicability of particular theories to diverse technologies in other fields. Thus, although this article aims to make a direct contribution to the area of autonomous driving ethics, it may also hold informative value for applied and machine ethics overall that raised similar questions in this regard (e.g., Berberich & Diepold, 2018). To validate this assumption, future studies should be conducted to thoroughly examine the generalizability of our results in other domains.

This article's findings are not equated with normative claims but can act as a basis for such future investigations. Although a summary of identified advantages and disadvantages of applying ethical theories to the decision-making of SDVs is provided, this article does not claim the moral superiority and invalidity of (integrating) particular theories nor the totality of all existing ethical theories at this stage. For example, it can be assumed that the identified literature here is skewed toward Western epistemic traditions and theories since much of the debates and investigations come from Euro-American scholars (Segun, 2021a). Therefore, the findings must be read and interpreted with caution. On a more general basis, this article aims to provide an understanding of the applicability of (a relatively comprehensive but not absolute list of) ethical theories to an SDV's decision-making from a functional perspective (i.e., what *could* be done?)

as well as point to corresponding social and moral/legal ramifications (i.e., what *would* be the consequences?). The moral legitimacy of implementing ethical theories into the decision-making of SDVs at all or in a particular manner (e.g., in the form of Figure 18) would need to be investigated in the future to generate a profound answer to what *should* be done. Nevertheless, the findings of this literature review can serve as a basis and reference point for this investigation. For example, looking at the blank boxes within the advantages/disadvantages matrices (Table 4 and Table 5), potential research gaps and underexplored research areas can be derived and addressed in future investigations. Furthermore, academics could explore to what extent the different evaluation criteria (social, moral/legal, and functional) are to be weighted equally or ranked in a particular order. For example, overemphasizing the functional perspective when programming SDVs may lead to technological solutionism in the sense that ethical theories that exhibit lower levels of technical feasibility are at risk of being neglected (Häußermann & Lütge, 2022). Such investigations – based on the insights of this article – can serve as useful in determining normative claims about the ethical decision-making of SDVs in the future.

There are many open questions and concrete figures to be determined to operationalize ethical decision-making in SDVs. As previously stated, Figure 18 summarizes how to integrate various ethical theories into SDV's decision-making. Due to the nature/methodology of this study, this model cannot specify more detailed instructions or numerical figures for the different layers. For example, this article does not define the numerical figure for a particular theory's weight/relative importance within an SDV's cost function. These specifications and quantifications, however, are key to direct SDVs' decision-making in the first place. Therefore, some open questions up for debate and future (empirical) investigations are, for example:

- What is the chronological order in which an SDV would apply particular theories during its decision-making process?
- Which theories, principles, and utilities should be considered and weighted in the SDV's cost function?
- If particular theories are integrated into an SDV's decision-making, to what extent and what kind of conflicts and tradeoffs will emerge between themselves?
- What are strict duties/rules that SDVs are not allowed to disobey? Moreover, what is the hierarchical order of these rules?
- What are legitimate valence factors for different traffic participants (e.g., pedestrians, cyclists, and vehicles), and how should SDVs weigh the different principles/theories in the cost function against each other?

- What are legitimate/necessary (risk) thresholds for SDVs? What are ‘reasonable’ factors for an SDV to take particular levels of risk?
- What, if any, are legitimate factors for arbitrary discrimination when it comes to traffic safety (e.g., targeting bigger vehicles since they are usually accompanied by higher levels of passenger safety; sparing an SDV’s passengers due to the manufacturer’s obligation towards their customers)?
- Should the ethical decision-making of SDVs change in different traffic scenarios? Namely, what are the corresponding utilities, rules, weights, thresholds, etc., depending on the typification of the situation?
- What metaethical considerations should direct the debates and developments in autonomous driving ethics? For example, should there be national or international standards/legislation for programming SDVs’ ethical decision-making?
- To what extent is it at all legitimate to operationalize ethical decision-making in SDVs?

Addressing technical issues may take precedence over the ethical programming of SDVs. As Siegel and Pappas (2021) state, “there remain ‘bigger fish to fry’ from a technical perspective, before such problems [i.e., the development of ethical code] warrant consideration” (p.8). After all, SDVs operate with reduced functionality in the presence of a critical fault in the vehicle (Németh, 2022). Similarly, when looking at the findings of this study, functional matters are essential for the instantiation of particular ethical theories in the first place. For example, full and precise information (e.g., of a collision’s outcomes) is needed for an SDV to be able to determine and select the trajectory that meets its duty to avoid injuring humans (i.e., exemplary instantiation of a deontological logic) or that minimizes personal injuries (i.e., exemplary instantiation of a utilitarian logic). However, state-of-the-art software and hardware of SDVs do not yet meet the required level of sophistication for a smooth operation or the consideration of certain factors (e.g., error-prone sensors or insufficient quantity and quality of training data) (Roff, 2018). Therefore, developing the hardware or finding solutions to limited data availability or poor data quality may present more immediate challenges, whereby close cooperation between philosophers and engineers can serve as a vehicle to “minimize the ‘gap’ between desired and feasible states” (Siegel & Pappas, 2021; p.10).

3.5 Conclusion

As SDVs will make decisions that hold ethical dimensions and manufacturers will need to predetermine these, it becomes key to evaluate the suitability and applicability of different ethical theories that could be drawn upon in this programming. Therefore, our review structures and integrates autonomous driving ethics literature by providing an overview of the social, moral/legal, and functional advantages and disadvantages of applying ethical theories to the decision-making of SDVs. This analysis aims to generate a more profound understanding of the possibilities and ramifications of consulting (i.e., programming) particular theories. Furthermore, we derive a model that shows, on a macro-level, that every identified ethical theory could be technically implemented in algorithms and guide SDVs' ethical decision-making. Overall, the purpose of this article is not to defend any single best theory but to set the groundwork for reflected integration and programming of ethical decision-making within SDVs and – to some extent – within machines overall. For automotive companies, the results of this review can enable them to develop a reflected value register/list, which may demonstrate their “forward-looking responsibility” (Nyholm & Smids, 2016; p.1283) and yield legal safeguarding in case the programming of a company's SDVs is questioned retrospectively. For policymakers, this review can help them comprehend which tradeoffs and negative consequences can be expected from SDVs that follow particular ethical theories, thereby pointing to the development of regulatory measures and restrictions. In the future, next to ethical discussions (e.g., investigation of the here posed open questions), addressing technical issues will serve as a gatekeeper to translate ethical considerations into code in the first place.

References

- Akrich, M. (1992). The De-Description of Technical Objects. In W. E. Bijker, & J. Law (Eds.), *Shaping Technology/Building Society*. The MIT Press.
- Alexander, L., & Moore, M. (2007). Deontological ethics. *The Stanford Encyclopedia of Philosophy*. Retrieved from: <https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=ethics-deontological>
- Altay, B. C., Boztas, A. E., Okumuş, A., Gul, M., & Çelik, E. (2023). How Will Autonomous Vehicles Decide in Case of an Accident? An Interval Type-2 Fuzzy Best–Worst Method for Weighting the Criteria from Moral Values Point of View. *Sustainability*, 15(11), 8916. <https://doi.org/10.3390/su15118916> [1]
- Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI magazine*, 28(4), 15-15. <https://doi.org/10.1609/aimag.v28i4.2065>
- Arfini, S., Spinelli, D., & Chiffi, D. (2022). Ethics of self-driving cars: A naturalistic approach. *Minds and Machines*, 32(4), 717-734. <https://doi.org/10.1007/s11023-022-09604-y> [2]
- Ashford, E., & Mulgan, T. (2018). Contractualism. *The Stanford Encyclopedia of Philosophy*. Retrieved from: <https://plato.stanford.edu/archives/sum2018/entries/contractualism/>
- Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature*, 563(7729), 59-64. <https://doi.org/10.1038/s41586-018-0637-6> [3]
- Bartneck, C., Lütge, C., Wagner, A., & Welsh, S. (2021). *An introduction to ethics in robotics and AI*. Springer nature. <https://doi.org/10.1007/978-3-030-51110-4>
- Basl, J., & Behrends, J. (2020). Why everyone has it wrong about the ethics of autonomous vehicles. In *Frontiers of Engineering: Reports on Leading-Edge Engineering from the 2019 Symposium*. National Academies Press. [4]
- Bennett, S. (2022). *Algorithms of life and death: a utilitarian approach to the ethics of self-driving cars*. Macquarie University. [5]
- Berberich, N., & Diepold, K. (2018). The Virtuous Machine-Old Ethics for New Technology?. *arXiv preprint*. <https://doi.org/10.48550/arXiv.1806.10322>
- Bergmann, L. T. (2022). Ethical issues in automated driving—Opportunities, dangers, and obligations. In A. Riener, M. Jeon, & I. Alvarez (Eds.), *User Experience Design in the Era of Automated Driving* (pp.99-121). Springer. https://doi.org/10.1007/978-3-030-77726-5_5
- Bergmann, L. T., Schlicht, L., Meixner, C., König, P., Pipa, G., Boshammer, S., & Stephan, A. (2018). Autonomous vehicles require socio-political acceptance – an empirical and philosophical perspective on the problem of moral decision making. *Frontiers in behavioral neuroscience*, 12, 31. <https://doi.org/10.3389/fnbeh.2018.00031> [6]
- Berkey, B. (forthcoming). Autonomous Vehicles, Business Ethics, and Risk Distribution in Hybrid Traffic. In *Autonomous Vehicle Ethics: Beyond the Trolley Problem*. [7]
- Bhargava, V., & Kim, T. W. (2017). Autonomous vehicles and moral uncertainty. In P. Lin, R. Jenkins, & K. Abney (Eds.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence* (pp.5-19). Oxford University Press. <https://doi.org/10.1093/oso/9780190652951.003.0001> [8]

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, 352(6293), 1573-1576. <https://doi.org/10.1126/science.aaf2654> [9]

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2019). The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars [point of view]. *Proceedings of the IEEE*, 107(3), 502-504. <https://doi.org/10.1109/JPROC.2019.2897447> [10]

Brandão, M. (2018). Moral autonomy and equality of opportunity for algorithms in autonomous vehicles. In M. Coeckelbergh, J. Loh, M. Funk, J. Seibt, & M. Nørskov (Eds.), *Envisioning Robots in Society—Power, Politics, and Public Space* (pp.302-310). IOS Press. <https://doi.org/10.3233/978-1-61499-931-7-302> [11]

Brändle, C., & Schmidt, M. W. (2021). Autonomous Driving and Public Reason: a Rawlsian Approach. *Philosophy & Technology*, 34(4), 1475-1499. <https://doi.org/10.1007/s13347-021-00468-1> [12]

Corley, K. G., & Gioia, D. A. (2004). Identity ambiguity and change in the wake of a corporate spin-off. *Administrative science quarterly*, 49(2), 173-208. <https://doi.org/10.2307/4131471>

Davnull, R. (2020). Solving the single-vehicle self-driving car trolley problem using risk theory and vehicle dynamics. *Science and Engineering Ethics*, 26(1), 431-449. <https://doi.org/10.1007/s11948-019-00102-6> [13]

DeLapp, K. M. (2011). Metaethics. *Internet Encyclopedia of Philosophy*. Retrieved from: <https://iep.utm.edu/metaethi/>

Deemantha, R. G. S., & Hettige, B (2022). *Autonomous Car: Current Issues, Challenges and Solution: A Review*. Retrieved from: https://www.researchgate.net/profile/Sasika-Deemantha/publication/366986201_Autonomous_Car_Current_Issues_Challenges_and_Solution_A_Review/links/63bd1296097c7832caa4fc62/Autonomous-Car-Current-Issues-Challenges-and-Solution-A-Review.pdf

De Melo, C. M., Marsella, S., & Gratch, J. (2021). Risk of injury in moral dilemmas with autonomous vehicles. *Frontiers in Robotics and AI*, 7, 572529. <https://doi.org/10.3389/frobt.2020.572529> [14]

Dietrich, M., & Weisswange, T. H. (2019). Distributive justice as an ethical principle for autonomous vehicle behavior beyond hazard scenarios. *Ethics and Information Technology*, 21(3), 227-239. <https://doi.org/10.1007/s10676-019-09504-3> [15]

Dogan, E., Chatila, R., Chauvier, S., Evans, K., Hadjixenophontos, P., & Perrin, J. (2016). Ethics in the Design of Automated Vehicles: The AVEthics project. In *EDIA@ ECAI* (pp.10-13). [16]

Dogan, E., Costantini, F., & Le Boennec, R. (2020). Ethical issues concerning automated vehicles and their implications for transport. *Advances in Transport Policy and Planning*, 5, 215-233. <https://doi.org/10.1016/bs.atpp.2020.05.003> [17]

Dolgoy, D. and Urmson, C. (2014). *Controlling vehicle lateral lane positioning*. Retrieved from: <https://patents.google.com/patent/US8781670B2/en>

D'Souza, J., Burnham, K. J., & Pickering, J. E. (2022). Modelling and Simulation of an Autonomous Vehicle Ethical Steering Control System (ESCS). In *2022 26th International*

Conference on Methods and Models in Automation and Robotics (MMAR) (pp.76-80). IEEE. <https://doi.org/10.1109/MMAR55195.2022.9874320> [18]

Dubljević, V. (2020). Toward implementing the ADC model of moral judgment in autonomous vehicles. *Science and Engineering Ethics*, 26(5), 2461-2472. <https://doi.org/10.1007/s11948-020-00242-0> [19]

Duncan, G. (2022). Deep Learning-based Ethical Judgments in Connected Vehicle Technologies: Route Planning Algorithms, Spatial Data Visualization Tools, and Real-Time Predictive Analytics. *Contemporary Readings in Law and Social Justice*, 14(2), 46-63.

Dyoub, A., Costantini, S., & Lisi, F. A. (2020). Logic programming and machine ethics. *arXiv preprint arXiv:2009.11186*. <https://doi.org/10.48550/arXiv.2009.11186>

Ebina, T., & Kinjo, K. (2021). Approaching the social dilemma of autonomous vehicles with a general social welfare function. *Engineering Applications of Artificial Intelligence*, 104, 104390. <https://doi.org/10.1016/j.engappai.2021.104390> [20]

Etienne, H. (2022). A practical role-based approach for autonomous vehicle moral dilemmas. *Big Data & Society*, 9(2), 20539517221123305. <https://doi.org/10.1177/2053951722112333> [21]

Etzioni, A., & Etzioni, O. (2017). Incorporating Ethics into Artificial Intelligence. *The Journal of Ethics*, 21(4), 403-418. <https://doi.org/10.1007/s10892-017-9252-2> [22]

European Commission (2020). *Ethics of connected and automated vehicles: Recommendations on road safety, privacy, fairness, explainability and responsibility*. Retrieved from: <https://op.europa.eu/en/publication-detail/-/publication/89624e2c-f98c-11ea-b44f-01aa75ed71a1/language-en/format-PDF/source-search>

European Commission (2021). *Regulatory framework proposal on artificial intelligence*. Retrieved from: <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>

Evans, K., de Moura, N., Chauvier, S., Chatila, R., & Dogan, E. (2020). Ethical decision making in autonomous vehicles: The AV ethics project. *Science and Engineering Ethics*, 26(6), 3285-3312. <https://doi.org/10.1007/s11948-020-00272-8> [23]

Faulhaber, A. K., Dittmer, A., Blind, F., Wächter, M. A., Timm, S., Sütthfeld, L. R., Stephan, A., Pipa, G., & König, P. (2019). Human decisions in moral dilemmas are largely described by utilitarianism: Virtual car driving study provides guidelines for autonomous driving vehicles. *Science and Engineering Ethics*, 25(2), 399-418. <https://doi.org/10.1007/s11948-018-0020-x> [24]

Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International journal of qualitative methods*, 5(1), 80-92. <https://doi.org/10.1177/16094069060050010>

Fossa, F. (2023). Unavoidable Collisions. The Automation of Moral Judgment. In F. Fossa (Ed.), *Ethics of Driving Automation: Artificial Agency and Human Values* (pp.65-94). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-22982-4_4 [25]

Fournier, T. (2016). Will my next car be a libertarian or a utilitarian?: Who will decide?. *IEEE Technology and Society Magazine*, 35(2), 40-45. <https://doi.org/10.1109/MTS.2016.2554441> [26]

Frank, D. A., Chrysochou, P., Mitkidis, P., & Ariely, D. (2019). Human decision-making biases in the moral dilemmas of autonomous vehicles. *Scientific reports*, 9(1), 1-19. <https://doi.org/10.1038/s41598-019-49411-7> [27]

Frison, A. K., Wintersberger, P., & Riemer, A. (2016). First person trolley problem: Evaluation of drivers' ethical decisions in a driving simulator. In *Adjunct proceedings of the 8th international conference on automotive user interfaces and interactive vehicular applications* (pp.117-122). <https://doi.org/10.1145/3004323.3004336> [28]

Geisslinger, M., Poszler, F., Betz, J., Lütge, C., & Lienkamp, M. (2021). Autonomous driving ethics: From trolley problem to ethics of risk. *Philosophy & Technology*, 34(4), 1033-1055. <https://doi.org/10.1007/s13347-021-00449-4> [29]

Geisslinger, M., Poszler, F., & Lienkamp, M. (2023). An ethical trajectory planning algorithm for autonomous vehicles. *Nature Machine Intelligence*, 5(2), 137-144. <https://doi.org/10.1038/s42256-022-00607-z> [30]

Gentzel, M. (2020). Classical liberalism, discrimination, and the problem of autonomous cars. *Science and Engineering Ethics*, 26(2), 931-946. <https://doi.org/10.1007/s11948-019-00155-7> [31]

Gerdes, J. C. (2020). The Virtues of Automated Vehicle Safety-Mapping Vehicle Safety Approaches to Their Underlying Ethical Frameworks. In *2020 IEEE Intelligent Vehicles Symposium (IV)* (pp.107-113). <https://doi.org/10.1109/IV47402.2020.9304583> [32]

Gerdes, J. C., & Thornton, S. M. (2015). Implementable ethics for autonomous vehicles. In M. Maurer, J. C. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomes Fahren – technische, rechtliche und gesellschaftliche Aspekte* (pp.87-102). Springer Vieweg. https://doi.org/10.1007/978-3-662-45854-9_5 [33]

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2012). Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology. *Organizational research methods*, 16 (1), 15–31. <https://doi.org/10.1177/1094428112452151>

Gogoll, J., & Müller, J. F. (2017). Autonomous cars: in favor of a mandatory ethics setting. *Science and Engineering Ethics*, 23(3), 681-700. <https://doi.org/10.1007/s11948-016-9806-x>

Goodall, N. J. (2014a). Ethical decision making during automated vehicle crashes. *Transportation Research Record: Journal of the Transportation Research Board*, 2424(1), 58–65. <https://doi.org/10.3141/2424-07> [34]

Goodall, N. J. (2014b). Machine ethics and automated vehicles. In G. Meyer & S. Beiker (Eds.), *Road vehicle automation* (pp.93–102). Springer International Publishing. https://doi.org/10.1007/978-3-319-05990-7_9 [35]

Goodall, N. J. (2016a). Away from trolley problems and toward risk management. *Applied Artificial Intelligence*, 30(8), 810-821. <https://doi.org/10.1080/08839514.2016.1229922> [36]

Goodall, N. J. (2016b). Can you program ethics into a self-driving car?. *IEEE Spectrum*, 53(6), 28-58. <https://doi.org/10.1109/MSPEC.2016.7473149> [37]

Goodall, N. (2019). More than trolleys: Plausible, ethically ambiguous scenarios likely to be encountered by automated vehicles. *Transfers*, 9(2), 45–58. <https://doi.org/10.3167/TRANS.2019.090204> [38]

Grasso, G. M., Lucifora, C., Perconti, P., & Plebe, A. (2020). Integrating human acceptable morality in autonomous vehicles. In T. Ahram, W. Karwowski, A. Vergnano, F. Leali, & R. Taiar (Eds.), *Intelligent Human Systems Integration* (pp.41-45). Springer. https://doi.org/10.1007/978-3-030-39512-4_7 [39]

Gurney, J. K. (2015). Crashing into the unknown: An examination of crash-optimization algorithms through the two lanes of ethics and law. *Albany Law Review*, 79(1), 183–267. [40]

Hansson, S. O., Belin, M. Å., & Lundgren, B. (2021). Self-driving vehicles—an ethical overview. *Philosophy & Technology*, 34(4), 1383-1408. <https://doi.org/10.1007/s13347-021-00464-5>

Häußermann, J. J., & Lütge, C. (2022). Community-in-the-loop: towards pluralistic value creation in AI, or—why AI needs business ethics. *AI and Ethics*, (2), 341-362. <https://doi.org/10.1007/s43681-021-00047-2>

Himmelreich, J. (2018). Never mind the trolley: The ethics of autonomous vehicles in mundane situations. *Ethical Theory and Moral Practice*, 21(3), 669-684. <https://doi.org/10.1007/s10677-018-9896-4> [41]

Holstein, T., Dodig-Crnkovic, G., & Pelliccione, P. (2021). Steps toward real-world ethics for self-driving cars: Beyond the trolley problem. In S. J. Thompson (Ed.), *Machine Law, Ethics, and Morality in the Age of Artificial Intelligence* (pp.85-107). IGI Global. <http://dx.doi.org/10.4018/978-1-7998-4894-3.ch006>

Hübner, D., & White, L. (2018). Crash algorithms for autonomous cars: How the trolley problem can move us beyond harm minimisation. *Ethical Theory and Moral Practice*, 21(3), 685-698. <https://doi.org/10.1007/s10677-018-9910-x> [42]

Hursthouse, R., & Pettigrove, G. (2016). Virtue Ethics. *The Stanford Encyclopedia of Philosophy*. Retrieved from: <https://plato.stanford.edu/entries/ethics-virtue/>

Islam, M. A., & Rashid, S. I. (2018). Algorithm for Ethical Decision Making at Times of Accidents for Autonomous Vehicles. In *2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEICT)* (pp.438-442). <https://doi.org/10.1109/CEEICT.2018.8628155> [43]

Jenkins, R. (2016). *Autonomous vehicles ethics & law*. Retrieved from: <https://d1y8sb8igg2f8e.cloudfront.net/documents/AV-Ethics-Law.pdf> [44]

Jiang, L., Xie, Y., & Evans, N. G. (2023). A simulation study of cooperative and autonomous vehicles (CAV) considering courtesy, ethics, and fairness. *Plos one*, 18(5), e0283649. <https://doi.org/10.1371/journal.pone.0283649> [45]

Johansson, R., & Nilsson, J. (2016). Disarming the trolley problem—why self-driving cars do not need to choose whom to kill. In *Workshop CARS 2016-Critical Automotive applications: Robustness & Safety*. [46]

Johnsen, A., Strand, N., Andersson, J., Patten, C., Kraetsch, C., & Takman, J. (2017). *Literature review on the acceptance and road safety, ethical, legal, social and economic implications of automated vehicles*. Institut für Empirische Soziologie an der Universität Erlangen-Nürnberg.

Karnouskos, S. (2018). Self-driving car acceptance and the role of ethics. *IEEE Transactions on Engineering Management*, 67(2), 252-265. <https://doi.org/10.1109/TEM.2018.2877307> [47]

Keeling, G. (2017). Against Leben's Rawlsian collision algorithm for autonomous vehicles. In *3rd Conference on Philosophy and Theory of Artificial Intelligence* (pp.259-272). Springer. https://doi.org/10.1007/978-3-319-96448-5_29 [48]

Keeling, G. (2018). Legal necessity, pareto efficiency & justified killing in autonomous vehicle collisions. *Ethical Theory and Moral Practice*, 21(2), 413-427. <https://doi.org/10.1007/s10677-018-9887-5> [49]

Keeling, G. (2020). Why trolley problems matter for the ethics of automated vehicles. *Science and Engineering Ethics*, 26(1), 293-307. <https://doi.org/10.1007/s11948-019-00096-1> [50]

Keeling, G., Evans, K., Thornton, S. M., Mecacci, G., & Santoni de Sio, F. (2018). Four perspectives on what matters for the ethics of automated vehicles. In G. Meyer, & S. Beiker (Eds.), *Road Vehicle Automation 6* (pp.49-60). Springer. https://doi.org/10.1007/978-3-030-22933-7_6 [51]

Kirchmair, L., & Paulo, N. (2023). Taking ethics seriously in AV trajectory planning algorithms. *Nature Machine Intelligence*, 5(8), 814-815. <https://doi.org/10.1038/s42256-023-00706-5>

Korosec, K. (2019). Waymo to customers: 'Completely driverless Waymo cars are on the way'. *TechCrunch*. Retrieved from: <https://techcrunch.com/2019/10/09/waymo-to-customers-completely-driverless-waymo-cars-are-on-the-way/>

Kriebitz, A., Max, R., & Lütge, C. (2022). The German Act on Autonomous Driving: why ethics still matters. *Philosophy & Technology*, 35(2), 1-13. <https://doi.org/10.1007/s13347-022-00526-2>

Krügel, S., & Uhl, M. (2022). Autonomous vehicles and moral judgments under risk. *Transportation research part A: policy and practice*, 155, 1-10. <https://doi.org/10.1016/j.tra.2021.10.016> [52]

Kumfer, W., & Burgess, R. (2015). Investigation into the role of rational ethics in crashes of automated vehicles. *Transportation research record*, 2489(1), 130-136. <https://doi.org/10.3141/2489-15> [53]

Latour, B., & Venn, C. (2002). Morality and technology. *Theory, culture & society*, 19(5-6), 247-260. <https://doi.org/10.1177/026327602761899246>

Leben, D. (2017). A Rawlsian algorithm for autonomous vehicles. *Ethics and Information Technology*, 19(2), 107-115. <https://doi.org/10.1007/s10676-017-9419-3> [54]

Leung, L. (2015). Validity, reliability, and generalizability in qualitative research. *Journal of family medicine and primary care*, 4(3), 324. <https://doi.org/10.4103/2249-4863.161306>

Li, J., Zhao, X., Cho, M. J., Ju, W., & Malle, B. F. (2016). From trolley to autonomous vehicle: Perceptions of responsibility and moral norms in traffic accidents with self-driving cars. *SAE Technical paper*, 10, 2016-01. <https://doi.org/10.4271/2016-01-0164> [55]

Li, L., Zhang, J., Wang, S., & Zhou, Q. (2022). A Study of Common Principles for Decision-Making in Moral Dilemmas for Autonomous Vehicles. *Behavioral Sciences*, 12(9), 344. <https://doi.org/10.3390/bs12090344> [56]

Lin, P. (2016). Why ethics matters for autonomous cars. In M. Maurer, J. C. Gerdes, B. Lenz & H. Winner (Eds.), *Autonomous driving: Technical, legal and social aspects* (pp.69–85). Springer. <https://doi.org/10.1007/978-3-662-48847-8> [57]

Liu, H. Y. (2017). Irresponsibilities, inequalities and injustice for autonomous vehicles. *Ethics and Information Technology*, 19(3), 193-207. <https://doi.org/10.1007/s10676-017-9436-2> [58]

Liu, P., & Liu, J. (2021). Selfish or utilitarian automated vehicles? Deontological evaluation and public acceptance. *International Journal of Human–Computer Interaction*, 37(13), 1231-1242. <https://doi.org/10.1080/10447318.2021.1876357> [59]

Lucifora, C., Grasso, G. M., Perconti, P., & Plebe, A. (2020). Moral dilemmas in self-driving cars. *Rivista internazionale di Filosofia e Psicologia*, 11(2), 238-250. <https://doi.org/10.4453/rifp.2020.0015> [60]

Lucifora, C., Grasso, G. M., Perconti, P., & Plebe, A. (2021). Moral reasoning and automatic risk reaction during driving. *Cognition, Technology & Work*, 23(4), 705-713. <https://doi.org/10.1007/s10111-021-00675-y> [61]

Lütge, C. (2017). The German ethics code for automated and connected driving. *Philosophy & Technology*, 30(4), 547-558. <https://doi.org/10.1007/s13347-017-0284-0> [62]

Lütge, C., Poszler, F., Acosta, A. J., Danks, D., Gottehrer, G., Mihet-Popa, L., & Naseer, A. (2021). AI4People: Ethical Guidelines for the Automotive Sector–Fundamental Requirements and Practical Recommendations. *International Journal of Technoethics*, 12(1), 101-125. <https://doi.org/10.4018/IJT.20210101.0a2>

Martinho, A., Herber, N., Kroesen, M., & Chorus, C. (2021). Ethical issues in focus by the autonomous vehicles industry. *Transport reviews*, 41(5), 556-577. <https://doi.org/10.1080/01441647.2020.1862355>

Mason, E. (2018). Value Pluralism. *The Stanford Encyclopedia of Philosophy*. Retrieved from: <https://plato.stanford.edu/entries/value-pluralism/#:~:text=Many%20utilitarians%20are%20monists%2C%20arguing,utilitarians%20are%20committed%20to%20hedonism.>

Mayer, M. M., Bell, R., & Buchner, A. (2021). Self-protective and self-sacrificing preferences of pedestrians and passengers in moral dilemmas involving autonomous vehicles. *PLoS one*, 16(12), e0261673. <https://doi.org/10.1371/journal.pone.0261673> [63]

Meder, B., Fleischhut, N., Krumnau, N. C., & Waldmann, M. R. (2019). How should autonomous cars drive? A preference for defaults in moral judgments under risk and uncertainty. *Risk analysis*, 39(2), 295-314. <https://doi.org/10.1111/risa.13178> [64]

Millán-Blanquel, L., Veres, S. M., & Purshouse, R. C. (2020). Ethical Considerations for a Decision Making System for Autonomous Vehicles During an Inevitable Collision. In *28th Mediterranean Conference on Control and Automation (MED)* (pp.514-519). <https://doi.org/10.1109/MED48518.2020.9183263> [65]

Mirrig, A. G., & Meschtscherjakov, A. (2019). Trolled by the trolley problem: on what matters for ethical decision making in automated vehicles. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp.1-10). <https://doi.org/10.1145/3290605.3300739> [66]

- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group, T. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine*, 151(4), 264-269. <https://doi.org/10.7326/0003-4819-151-4-200908180-00135>
- Motwani, S., Sharma, T., & Gupta, A. (2021). Ethics in Autonomous Vehicle Software: The Dilemmas. *Computer*, 54(8), 46-55. <https://doi.org/10.1109/MC.2021.3077576> [67]
- Narayanan, A. (2019). Ethical judgement in intelligent control systems for autonomous vehicles. In 2019 *Australian & New Zealand Control Conference (ANZCC)* (pp.231-236). <https://doi.org/10.1109/ANZCC47194.2019.8945790> [68]
- Nath, R., & Sahu, V. (2020). The problem of machine ethics in artificial intelligence. *AI & society*, 35(1), 103-111. <https://doi.org/10.1007/s00146-017-0768-6>
- Németh, B. (2022). Route selection method with ethical considerations for automated vehicles under critical situations. In 2022 *IEEE 20th Jubilee World Symposium on Applied Machine Intelligence and Informatics (SAMI)* (pp.000419-000424). <https://doi.org/10.1109/SAMI54271.2022.9780742> [69]
- Németh, B. (2023). Coordinated Control Design for Ethical Maneuvering of Autonomous Vehicles. *Energies*, 16(10), 4254. <https://doi.org/10.3390/en16104254> [70]
- Nida-Rümelin, J., Schulenburg, J., & Rath, B. (2012). *Risikoethik*. Walter de Gruyter.
- Novak, T. P. (2020). A generalized framework for moral dilemmas involving autonomous vehicles: a commentary on gill. *Journal of Consumer Research*, 47(2), 292-300. <https://doi.org/10.1093/jcr/ucaa024> [71]
- Nyholm, S. (2018). The ethics of crashes with self-driving cars: A roadmap, I. *Philosophy Compass*, 13(7), e12507. <https://doi.org/10.1111/phc3.12507> [72]
- Nyholm, S. (2023). *This is Technology Ethics: An introduction*. John Wiley & Sons.
- Nyholm, S., & Smids, J. (2016). The ethics of accident-algorithms for self-driving cars: An applied trolley problem?. *Ethical theory and Moral Practice*, 19(5), 1275-1289. <https://doi.org/10.1007/s10677-016-9745-2> [73]
- Pan, S., Thornton, S. M., & Gerdes, J. C. (2016). Prescriptive and proscriptive moral regulation for autonomous vehicles in approach and avoidance. In 2016 *IEEE International Symposium on Ethics in Engineering, Science and Technology (ETHICS)* (pp.1-6). <https://doi.org/10.1109/ETHICS.2016.7560049> [74]
- Papadimitriou, E., Farah, H., van de Kaa, G., De Sio, F. S., Hagenzieker, M., & van Gelder, P. (2022). Towards common ethical and safe 'behaviour' standards for automated vehicles. *Accident Analysis & Prevention*, 174, 106724. <https://doi.org/10.1016/j.aap.2022.106724>
- Pickering, J. E., Podsiadly, M., & Burnham, K. J. (2019). A model-to-decision approach for the autonomous vehicle (av) ethical dilemma: Av collision with a barrier/pedestrian (s). *IFAC-PapersOnLine*, 52(8), 257-264. <https://doi.org/10.1016/j.ifacol.2019.08.080> [75]
- Pickering, J., & D'Souza, J. (2023). *Deontological Ethics for Safe and Ethical Algorithms for Navigation of Autonomous Vehicles (C-NAV) on a Highway*. Retrieved from: https://www.researchgate.net/profile/James-Pickering-2/publication/370984665_Deontological_Ethics_for_Safe_and_Ethical_Algorithms_for_Navigation_of_Autonomous_Vehicles_C-NAV_on_a_Highway/links/659000000000000000000000/Deontological-Ethics-for-Safe-and-Ethical-Algorithms-for-Navigation-of-Autonomous-Vehicles-C-NAV-on-a-Highway.pdf

[NAV on a Highway/links/646dd48437d6625c002c81fe/Deontological-Ethics-for-Safe-and-Ethical-Algorithms-for-Navigation-of-Autonomous-Vehicles-C-NAV-on-a-Highway.pdf](#) [76]

Qian, Z., Guo, P., Wang, Y., & Xiao, F. Ethical and moral decision-making for self-driving cars based on deep reinforcement learning. *Journal of Intelligent & Fuzzy Systems*, (Preprint), 1-18. <https://doi.org/10.3233/JIFS-224553> [77]

Rafiee, A., Wu, Y., & Sattar, A. (2023). Philosophical and Legal Approach to Moral Settings in Autonomous Vehicles: An Evaluation. In H. Breakey (Ed.), *Social Licence and Ethical Practice* (pp.95-114). Emerald Publishing Limited. <https://doi.org/10.1108/S1529-209620230000027007> [78]

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

Reed, N., Leiman, T., Palade, P., Martens, M., & Kester, L. (2021). Ethics of automated vehicles: breaking traffic rules for road safety. *Ethics and Information Technology*, 23(4), 777-789. <https://doi.org/10.1007/s10676-021-09614-x> [79]

Rhim, J., Lee, G. B., & Lee, J. H. (2020). Human moral reasoning types in autonomous vehicle moral dilemma: a cross-cultural comparison of Korea and Canada. *Computers in Human Behavior*, 102, 39-56. <https://doi.org/10.1016/j.chb.2019.08.010> [80]

Robinson, J., Smyth, J., Woodman, R., & Donzella, V. (2021a). Ethical considerations and moral implications of autonomous vehicles and unavoidable collisions. *Theoretical Issues in Ergonomics Science*, 23(4), 435-452. <https://doi.org/10.1080/1463922X.2021.1978013> [81]

Robinson, P., Sun, L., Furey, H., Jenkins, R., Phillips, C. R., Powers, T. M., Ritterson, R. S., Xie, Y., Casagrande, R., & Evans, N. G. (2021b). Modelling Ethical Algorithms in Autonomous Vehicles Using Crash Data. *IEEE Transactions on Intelligent Transportation Systems*, 23(7), 7775-7784. <https://doi.org/10.1109/TITS.2021.3072792> [82]

Roff, H. (2018). The folly of trolleys: Ethical challenges and autonomous vehicles. *The Brookings Institution*. Retrieved from: <https://www.brookings.edu/research/the-folly-of-trolleys-ethical-challenges-and-autonomous-vehicles/> [83]

Ryazanov, A. A., Wang, S. T., Nelkin, D. K., McKenzie, C. R., & Rickless, S. C. (2023). Beyond killing one to save five: Sensitivity to ratio and probability in moral judgment. *Journal of Experimental Social Psychology*, 108, 104499. <https://doi.org/10.1016/j.jesp.2023.104499> [84]

SAE International (2018). *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*. Retrieved from: https://www.sae.org/standards/content/j3016_201806/

Sætra, H. S., & Danaher, J. (2022). To each technology its own ethics: The problem of ethical proliferation. *Philosophy & Technology*, 35(4), 93. <https://doi.org/10.1007/s13347-022-00591-7>

Samuel, S., Yahoodik, S., Yamani, Y., Valluru, K., & Fisher, D. L. (2020). Ethical decision making behind the wheel—a driving simulator study. *Transportation research interdisciplinary perspectives*, 5, 100147. <https://doi.org/10.1016/j.trip.2020.100147> [85]

Scanlon, T. M. (1998). *What We Owe to Each Other*. Harvard University Press.

Schäffner, V. (2018). Caught up in ethical dilemmas: an adapted consequentialist perspective on self-driving vehicles. In M. Coeckelbergh, J. Loh, M. Funk, J. Seibt, & M. Nørskov (Eds.), *Envisioning Robots in Society—Power, Politics, and Public Space* (pp.327-335). IOS Press. [86]

Segun, S. T. (2021a). Critically engaging the ethics of AI for a global audience. *Ethics and Information Technology*, 23(2), 99-105. <https://doi.org/10.1007/s10676-020-09570-y>

Segun, S. T. (2021b). From machine ethics to computational ethics. *AI & SOCIETY*, 36(1), 263-276. <https://doi.org/10.1007/s00146-020-01010-1>

Shea-Blymyer, C., & Abbas, H. (2021). Algorithmic Ethics: Formalization and Verification of Autonomous Vehicle Obligations. *ACM Transactions on Cyber-Physical Systems (TCPS)*, 5(4), 1-25. <https://doi.org/10.1145/3460975> [87]

Siegel, J., & Pappas, G. (2021). Morals, ethics, and the technology capabilities and limitations of automated and self-driving vehicles. *AI & SOCIETY*, 38(1), 213-226. <https://doi.org/10.1007/s00146-021-01277-y> [88]

Smith, B. (2019). Personality facets and ethics positions as directives for self-driving vehicles. *Technology in Society*, 57, 115-124. <https://doi.org/10.1016/j.techsoc.2018.12.006> [89]

Soh, E., & Martens, K. (2022). Value dimensions of autonomous vehicle implementation through the Ethical Delphi. *Cities*, 127, 103741. <https://doi.org/10.1016/j.cities.2022.103741>

State of California – Department of Motor Vehicles (2019). *Autonomous Vehicle Disengagement Reports 2018*. Retrieved from: https://www.dmv.ca.gov/portal/dmv/detail/vr/autonomous/disengagement_report_2018

Sütfeld, L. R., Gast, R., König, P., & Pipa, G. (2017). Using virtual reality to assess ethical decisions in road traffic scenarios: applicability of value-of-life-based models and influences of time pressure. *Frontiers in behavioral neuroscience*, 11, 122. <https://doi.org/10.3389/fnbeh.2017.00122> [90]

Sütfeld, L. R., König, P., & Pipa, G. (2019). Towards a framework for ethical decision making in automated vehicles. *PsyArXiv preprint*. <https://doi.org/10.31234/osf.io/4duca> [91]

Theurer, C. P., Tumasjan, A., Welp, I. M., & Lievens, F. (2018). Employer branding: a brand equity-based literature review and research agenda. *International Journal of Management Reviews*, 20(1), 155-179. <https://doi.org/10.1111/ijmr.12121>

Thomson, J. J. (1984). The trolley problem. *The Yale Law Journal*, 94, 1395, 1985.

Thornton, S. M., Pan, S., Erlien, S. M., & Gerdes, J. C. (2016). Incorporating ethical considerations into automated vehicle control. *IEEE Transactions on Intelligent Transportation Systems*, 18(6), 1429-1439. <https://doi.org/10.1109/TITS.2016.2609339> [92]

Trappl, R. (2015). Robots' Ethical Systems: From Asimov's Laws to Principlism, from Assistive Robots to Self-Driving Cars. In R. Trappl (Ed.), *A Construction Manual for Robots' Ethical Systems* (pp.1-8). Springer. https://doi.org/10.1007/978-3-319-21548-8_1 [93]

Verbeek, P. P. (2011). *Moralizing technology: Understanding and designing the morality of things*. University of Chicago press.

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Wang, H., Huang, Y., Khajepour, A., Cao, D., & Lv, C. (2020). Ethical decision-making platform in autonomous vehicles with lexicographic optimization based model predictive

controller. *IEEE transactions on vehicular technology*, 69(8), 8164-8175. <https://doi.org/10.1109/TVT.2020.2996954> [94]

Wang, H., Khajepour, A., Cao, D., & Liu, T. (2022a). Ethical decision making in autonomous vehicles: Challenges and research progress. *IEEE Intelligent Transportation Systems Magazine*, 14(1), 6-17. <https://doi.org/10.1109/MITS.2019.2953556> [95]

Wang, Y., Hu, X., Yang, L., & Huang, Z. (2022b). Ethics Dilemmas and Autonomous Vehicles: Ethics Preference Modeling and Implementation of Personal Ethics Setting for Autonomous Vehicles in Dilemmas. *IEEE Intelligent Transportation Systems Magazine*, 15(2), 177-189. <https://doi.org/10.1109/MITS.2022.3197689> [96]

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS quarterly*, 26(2), xiii-xxiii.

Wintersberger, P., Prison, A. K., Riener, A., & Hasirlioglu, S. (2017). The experience of ethics: Evaluation of self harm risks in automated vehicles. In *2017 IEEE Intelligent Vehicles Symposium (IV)* (pp.385-391). <https://doi.org/10.1109/IVS.2017.7995749> [97]

Wolkenstein, A. (2018). What has the Trolley Dilemma ever done for us (and what will it do in the future)? On some recent debates about the ethics of self-driving cars. *Ethics and Information Technology*, 20(3), 163-173. <https://doi.org/10.1007/s10676-018-9456-6> [98]

Woodgate, J. M., & Ajmeri, N. (2022). Macro Ethics for Governing Equitable Sociotechnical Systems. In *AAMAS '22: Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS)* (pp.1824-1828). IFAAMAS Press. <https://doi.org/10.5555/3535850.3536118>

Zhao, H., Dimovitz, K., Staveland, B., & Medsker, L. (2016). Responding to challenges in the design of moral autonomous vehicles. In *2016 AAAI Fall Symposium Series: Cognitive Assistance in Government and Public Sector Applications*. [99]

Zhao, L., & Li, W. (2020). “Choose for No Choose” – Random-Selecting Option for the Trolley Problem in Autonomous Driving. In *LISS2019 – Proceedings of the 9th International Conference on Logistics, Informatics and Service Sciences* (pp.665-672). Springer. https://doi.org/10.1007/978-981-15-5682-1_48 [100]

Zhu, A., Yang, S., Chen, Y., & Xing, C. (2022). A moral decision-making study of autonomous vehicles: Expertise predicts a preference for algorithms in dilemmas. *Personality and Individual Differences*, 186, 111356. <https://doi.org/10.1016/j.paid.2021.111356> [101]

Appendix A of Essay II – Literature search & analysis

Database	Link	Query	Date	Hits	Hits (without duplicates)	Hits (full text analysis)	Forward & backward search	Hits (total)	Comments
Ebscohost business source premier	https://web.s.ebscohost.com/ehost/search/advanced?vid=1&sid=e7bed7d4-85e6-4abf-822a-260535ec417b%40redis	TI (("ethic" OR "moral" OR "philosoph" OR "norm") AND ("self" driv" OR "self" pilot" OR "driverless" OR "automat" automobile" OR "automat" vehicle" OR "automat" car" OR "autonomous" automobile" OR "autonomous" vehicle" OR "autonomous" car"))	31.07.2023	95	715	77	24	101	Limiters Language: English
ScienceDirect	https://www.sciencedirect.com/search	Terms: ("self driving" OR "self pilot" OR "driverless" OR "automated automobile" OR "automated vehicle" OR "automated car" OR "autonomous automobile" OR "autonomous vehicle" OR "autonomous car") Title, abstract or author-specified keywords: ("ethics" OR "moral" OR "philosoph" OR "norm")		457					
scopus	https://www.scopus.com/	TITLE (("ethic" OR "moral" OR "philosoph" OR "norm") AND ("self" driv" OR "self" pilot" OR "driverless" OR "automat" automobile" OR "automat" vehicle" OR "automat" car" OR "autonomous" automobile" OR "autonomous" vehicle" OR "autonomous" car")) AND (LIMIT-TO (LANGUAGE , "English"))		215					
Web of Science	https://apps.webofknowledge.com/	TI=(("ethic" OR "moral" OR "philosoph" OR "norm") AND ("self" driv" OR "self" pilot" OR "driverless" OR "automat" automobile" OR "automat" vehicle" OR "automat" car" OR "autonomous" automobile" OR "autonomous" vehicle" OR "autonomous" car"))		117					Refined by: Langu AGES: (ENGLISH)
SUM:				884					

Table 7: Overview of the literature search process

The list of keywords underlying the database search was refined in consultation among the authors to avoid selection bias, i.e., overlooking terms/concepts relevant to the topic. As illustrated in Table 7, the initial literature search resulted in a total of 884 hits, respectively 715 hits without duplicates. After the title and abstract analysis, 77 journal articles, book chapters, conference proceedings, and white papers were considered relevant to the topic at hand and, thus, subjected to the full paper analysis. In addition to the search in the four stated databases, forward- and backward searches were conducted by reviewing the reference lists of the 77 initially identified publications and utilizing Google Scholar's "cited by" function. These forward- and backward searches revealed another 24 fitting contributions. Therefore, 101 publications were identified as the baseline for this review and, thus, subject to further analysis. Adapted from the PRISMA flowchart for study selection (Moher et al., 2009), Figure 19 itemizes the entire search funnel that yielded the final sample of the literature analyzed in this review.

In line with Leung (2015), triangulation among authors was conducted by repeatedly consulting and agreeing upon generated codes and themes. In addition, drafts of the working paper were presented and discussed with fellow researchers and practitioners at international conferences.

Essay II: Applying Ethical Theories to the Decision-Making of Self-Driving Vehicles: A Systematic Review and Integration of the Literature

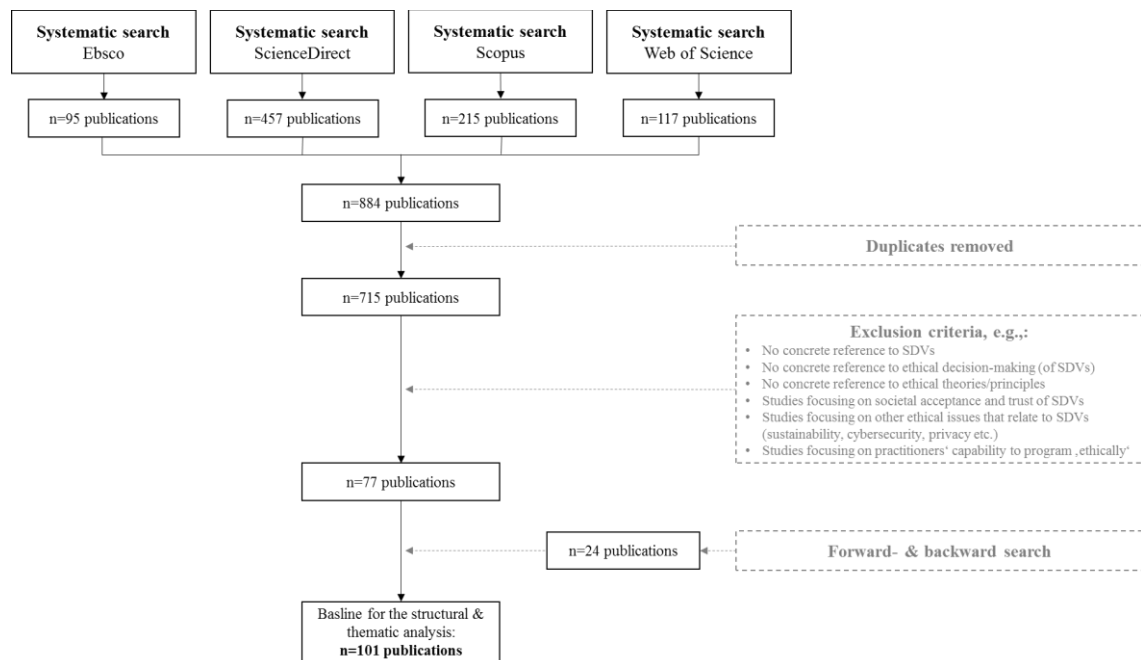


Figure 19: Search funnel of the systematic literature review, adapted from Moher et al. (2009)

Appendix B of Essay II – Descriptive/structural analysis of the literature

The structural analysis yields essential findings about the development of the automated driving ethics literature.

First, Figure 20 depicts a chronological overview (2014 to/including July 2023) of the identified publications by discipline. The first respective publications emerged in 2014. Since then, we have seen a steady increase in autonomous driving ethics publications. It is expected that the publication number for “2022-2023” will increase beyond 21, as it is likely that additional publications will be issued in the second half of the year. The overview also illustrates the multidisciplinary nature/plurality of disciplines of the autonomous driving ethics field, indicated by the assigned Web of Science category and/or Scopus subject area for the respective publication.

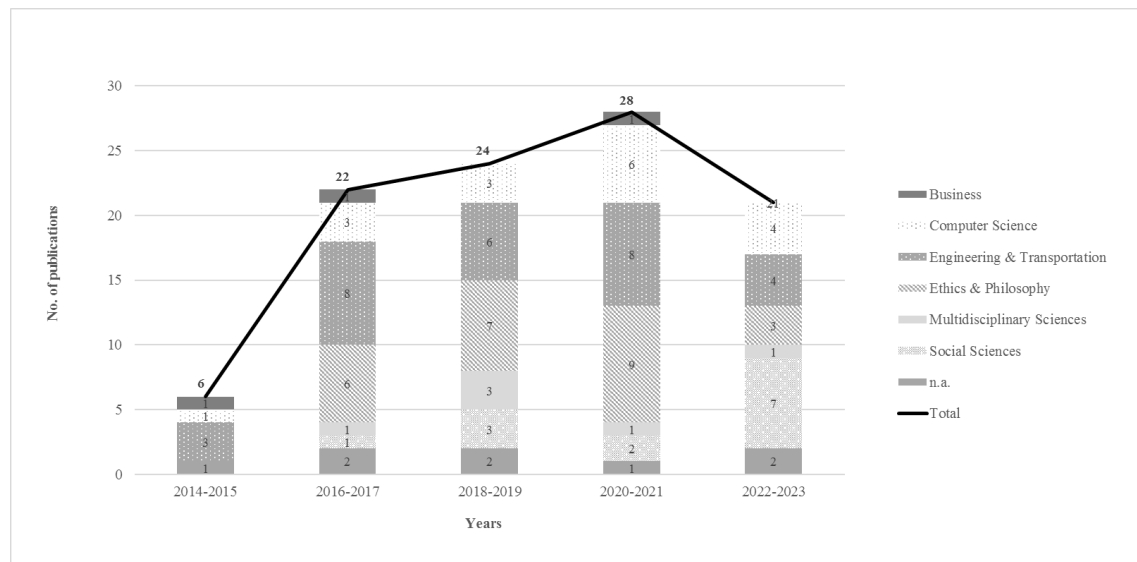


Figure 20: Automated driving ethics publications over time and by discipline (2014-July 2023)

Second, Table 8 provides an overview of all 101 publications of corresponding research outlets by discipline and journal/publication type. With a total number of 70, the majority of publications are journal articles (69.3%). Additionally, pertinent outlets are conference proceedings (21.8%), book chapters (7.9%), and one white paper (1%). As illustrated, most relevant publications were issued in the disciplines of Engineering & Transportation (28.7%) and Ethics & Philosophy (24.8%). Following this, the field of Computer Science ranks third with 17 (16.8%) pertinent publications. The discipline of Social Sciences issued 13 (12.9%) publications, and in the discipline of Multidisciplinary Sciences, 6 (5.9%) relevant publications were identified. Within the discipline of Business 3 (3%), corresponding publications were published. 8 (7.9%) publications were not classified according to the Web of Science category and/or Scopus subject area and are therefore clustered into ‘n.a.’.

Essay II: Applying Ethical Theories to the Decision-Making of Self-Driving Vehicles: A Systematic Review and Integration of the Literature

Primary subject category / discipline	Journal / publication type	No. of publications (%)	
Business	Book (chapter)	2	3 (3%)
	Journal of Consumer Research	1	
Computer Science	ACM Transactions on Cyber-Physical Systems	1	17 (16.8%)
	AI & Society	1	
	Applied Artificial Intelligence	1	
	Book (chapter)	1	
	Computer	1	
	Conference Proceeding	8	
	Frontiers in Robotics and AI	1	
	International Journal of Human-Computer Interaction	1	
	Minds and Machines	1	
	Nature Machine Intelligence	1	
Engineering & Transportation	Book (chapter)	1	29 (28.7%)
	Cognition, Technology & Work	1	
	Conference Proceeding	12	
	Energies	1	
	Engineering Applications of Artificial Intelligence	1	
	IEEE Intelligent Transportation Systems Magazine	2	
	IEEE Spectrum	1	
	IEEE Technology and Society Magazine	1	
	IEEE Transactions on Engineering Management	1	
	IEEE Transactions on Intelligent Transportation Systems	2	
	IEEE Transactions on Vehicular Technology	1	
	SAE Technical paper	1	
	Transportation Research Interdisciplinary Perspectives	1	
	Transportation Research Part A: Policy and Practice	1	
	Transportation Research Record	2	
Ethics & Philosophy	Book (chapter)	4	25 (24.8%)
	Ethical Theory and Moral Practice	4	
	Ethics and Information Technology	5	
	Philosophy & Technology	3	
	Philosophy Compass	1	
	Rivista internazionale di Filosofia e Psicologia	1	
	Science and Engineering Ethics	6	
	The Journal of Ethics	1	
Multidisciplinary Sciences	Nature	1	6 (5.9%)
	PLoS one	2	
	Risk analysis	1	
	Science	1	
	Scientific reports	1	
Social Sciences	Advances in Transport Policy and Planning	1	13 (12.9%)
	Behavioral Sciences	1	
	Big Data & Society	1	
	Computers in Human Behavior	1	
	Frontiers in Behavioral Neuroscience	2	
	Journal of Experimental Social Psychology	1	
	Personality and Individual Differences	1	
	Social Licence and Ethical Practice	1	
	Sustainability	1	
	Technology in Society	1	
	Theoretical Issues in Ergonomics Science	1	
	Transfers	1	
n.a.	Albany Law Review	1	8 (7.9%)
	Conference Proceeding	2	
	Journal of Intelligent & Fuzzy Systems	1	
	New america	1	
	PsyArXiv	1	
	The Brookings Institution	1	
	Whitepaper	1	
		Total:	101 (100%)

Table 8: Overview of research outlets by discipline (indicated by the assigned Web of Science category and/or Scopus subject area for the respective publication) and by journal/publication type

Third, Figure 21 depicts the research methods applied in the respective literature²⁰. In the past, conceptual (i.e., theoretical or mathematical) publications have dominated the research field, with, for example, over 18 publications in the years 2020-2021. 2022-2023 mark the first years in which conceptual articles are not the majority but make up only around 38%. Empirical studies that drew on simulations remained relatively constant between 2014 and 2021, issuing a maximum of 2 publications every two years. In 2022-2023, simulation studies sparked a number of 8 publications. Empirical studies that collect quantitative data (e.g., variations of Awad et al.'s (2018) 'Moral Machine experiment') have picked up and increased slowly since 2014, yielding 7 respective publications in 2022-2023.

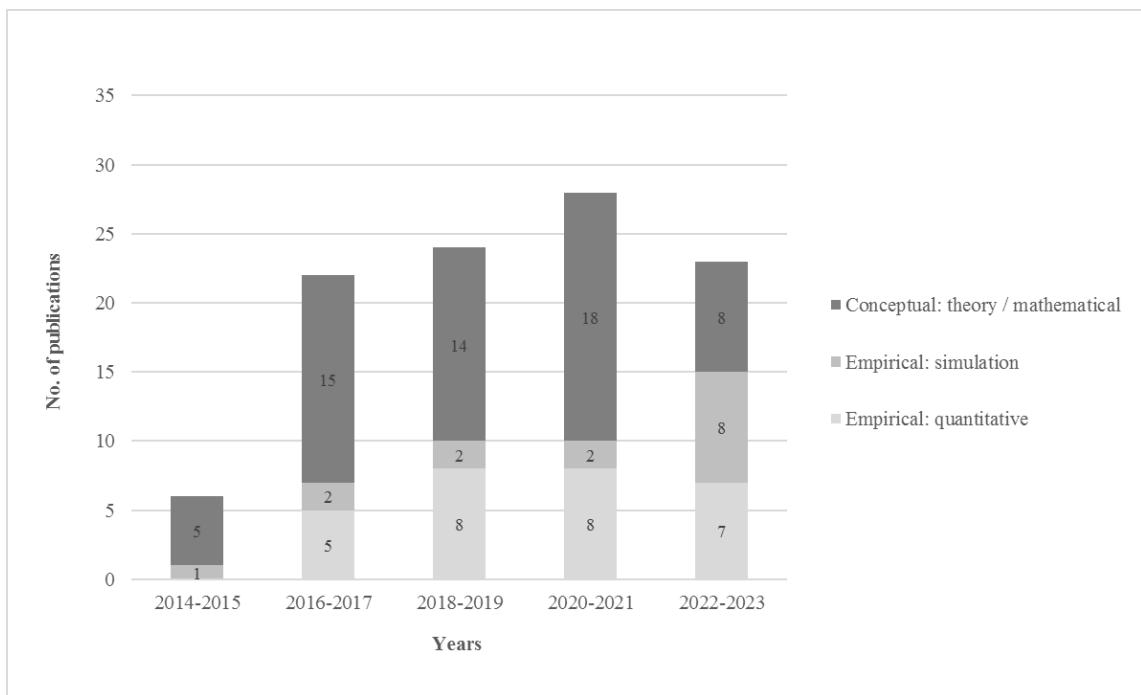


Figure 21: Utilized methodology of the identified publications

Fourth, we analyzed the practicability of the identified publications (see Figure 22). Overall, three degrees of practicability were recognized: (1) no remark on technical implementation (e.g., respective articles are rather argumentative concerning the implications of particular theories), (2) conceptualization of technical implementation (e.g., respective articles propose a conceptual model or framework of how to implement particular theories) or (3) concrete suggestion of technical implementation (e.g., respective articles provide a function or algorithm that is built on particular theories). The practicability of the publications was independently determined by two

²⁰ Note: Some articles utilized more than one method (e.g., authors of an article tested a decision-making model in a simulation and conducted a survey to assess its social acceptability) so that the indicated numbers in this figure may differ from the total number of publications stated in Figure 20.

of the authors. If the authors came to different ‘practicability’ categorizations, the publications were jointly re-assessed and discussed until the matching categorization was reached.

As illustrated in Figure 22, over time and still to this date, publications with no remark on the technical implementation make up the majority of the publications, with over 12 publications in the years 2022-2023. Nevertheless, publications that offer such low practicability are decreasing compared to the previous years, especially when considered in proportion: For example, between 2016 and 2021, such publications made up between 82% and 71%, while in 2022-2023, such publications made up 57%. Publications that conceptualize a technical implementation have remained low and relatively constant since 2014 (i.e., a maximum of 4 publications every two years). Publications that provide concrete suggestions of how to technically implement specific theories (e.g., in an algorithm) have steadily increased over time. Pertinent research has increased from 2 publications in 2014-2015 to 7 publications in 2022-2023. Exemplary studies are Geisslinger et al. (2023), Pickering and D’Souza (2023), and Németh (2023).

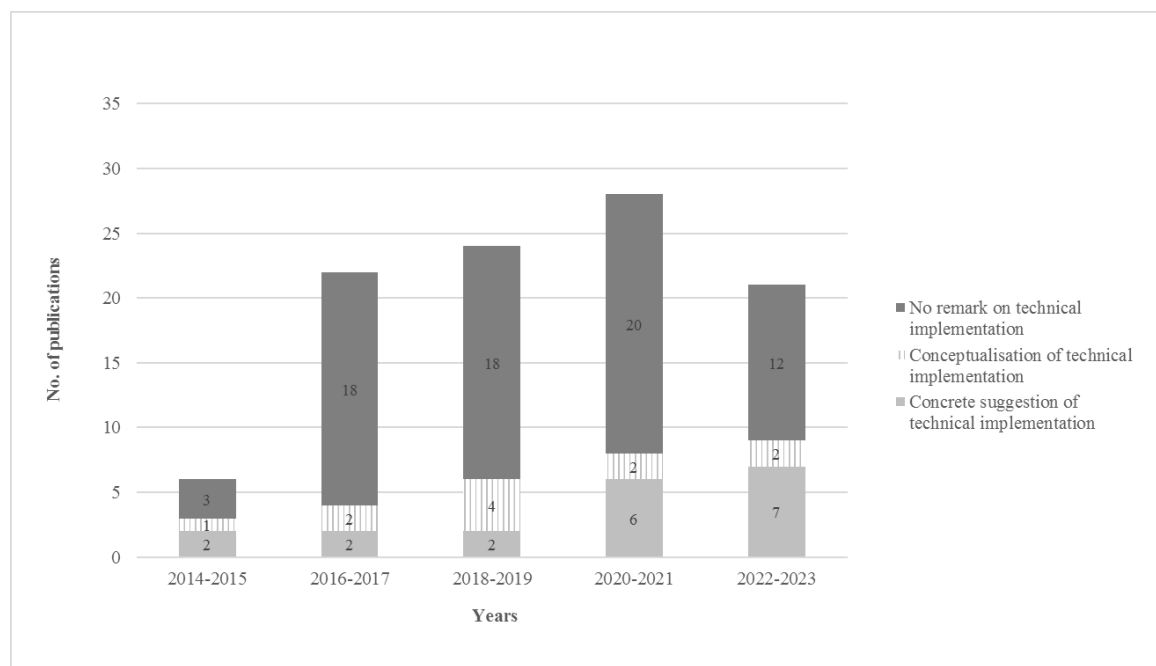


Figure 22: Practicability of the identified publications

Fifth, we analyzed the level of theory integration of the identified publications (see Figure 23). The publications were clustered into either ‘no theory integration’ (i.e., if no diverse theories are combined in proposed models/algorithms) or ‘theory integration’ (i.e., if proposed models/algorithms integrate two or more theories). An example of a publication clustered into ‘theory integration’ is the article by Gerdes and Thornton (2015), who combine deontological ethics/important moral rules as constraints (e.g., an imperative to avoid personal damage) and

utilitarianism in the form of weighing costs and options. Similar to categorizing a publication's practicability, the degree of theory integration was determined by and consulted between two authors.

Figure 23 illustrates the publications' degree of theory integration, classified by practicability. In general, publications with no theory integration make up most (71%), while only 29 (29%) publications conducted a theory integration. Over time, publications with theory integration have slowly increased from 3 publications in 2014-2015 to 9 in 2022-2023.

By classifying the publications according to the level of theory integration as well as practicability, it can be seen that they correlate with each other, i.e., publications conducted more theory integration with increasing practicability. For example, publications that provide no remark on technical implementation simultaneously do not integrate any theories. Among the publications that conceptualized a technical implementation, 10 integrated theories in their model/framework, while only one publication conducted no theory integration. Among the publications of the highest level of practicability (i.e., 'concrete suggestion of technical implementation'), all 19 publications showed theory integration.

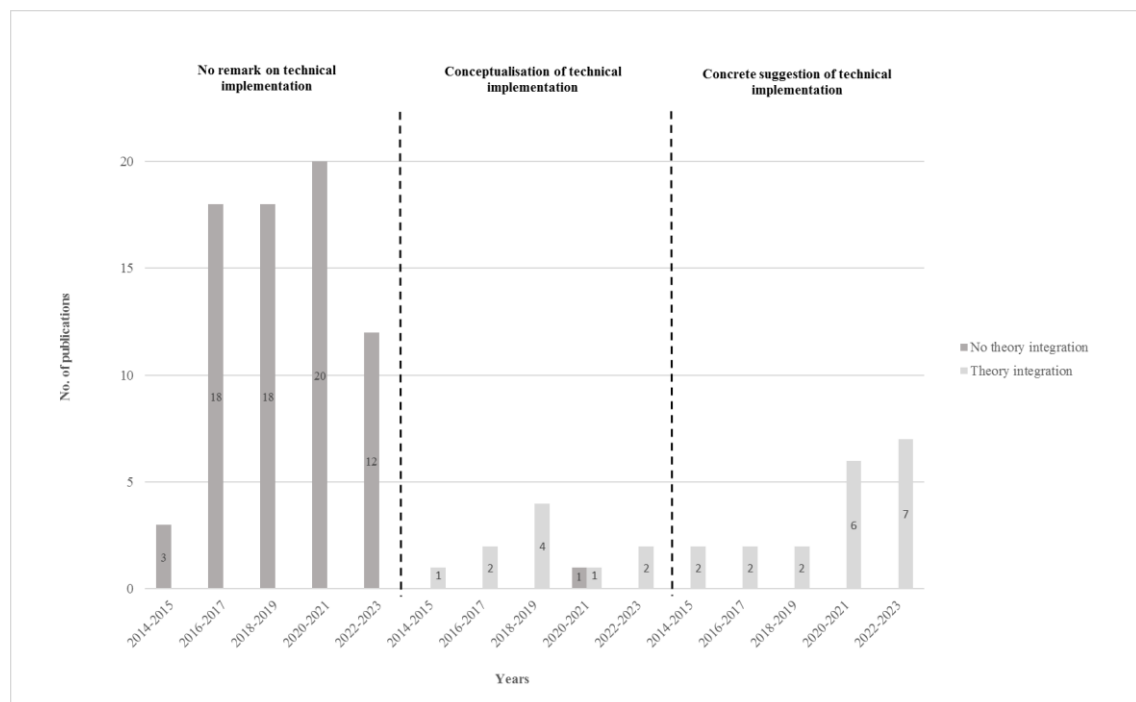


Figure 23: Theory integration classified by practicability of the identified publications

4 | Essay III: Ethical Decision-Making for Self-Driving Vehicles: A Proposed Model & List of Value-Laden Terms That Warrant (Technical) Specification

Abstract

Self-driving vehicles (SDVs) will need to make decisions that carry ethical dimensions and are of normative significance. For example, by choosing a specific trajectory, they determine how risks are distributed among traffic participants. Accordingly, policymakers, standardization organizations, and scholars have conceptualized what (shall) constitute(s) ethical decision-making for SDVs. Eventually, these conceptualizations must be converted into specific system requirements to ensure proper technical implementation. Therefore, this article aims to translate critical requirements recently formulated in scholarly work, existing standards, regulatory drafts, and guidelines into an explicit five-step ethical decision model for SDVs during hazardous situations. This model states a precise sequence of steps, indicates the guiding ethical principles that inform each step, and points out a list of terms that demand further investigation and technical specification. By integrating ethical, legal, and engineering considerations, we aim to contribute to the scholarly debate on computational ethics (particularly in autonomous driving) while offering practitioners in the automotive sector a decision-making process for SDVs that is technically viable, legally permissible, ethically grounded, and adaptable to societal values. In the future, assessing the actual impact, effectiveness, and admissibility of implementing the here-sketched theories, terms, and the overall decision process requires an empirical evaluation and testing of the overall decision-making model.

Keywords: *self-driving vehicle, autonomous driving, ethical decision-making, computational ethics, risk distributions*

Note: This chapter is based on a published paper co-authored by Maximilian Geisslinger and Christoph Lütge. Therefore, the plural instead of the singular is used throughout this chapter. Author contributions to this paper and copyright information are summarized in Appendix C: Reference & copyright information

Essay III: Ethical Decision-Making for Self-Driving Vehicles: A Proposed Model & List of Value-Laden Terms That Warrant (Technical) Specification

by the publisher for the third essay (Essay III, Chapter 4) and Appendix D: Author contributions to the three essays in this dissertation.

Current publication status:

Poszler, F., Geisslinger, M., & Lütge, C. (2024). Ethical Decision-Making for Self-Driving Vehicles: A Proposed Model & List of Value-Laden Terms that Warrant (Technical) Specification. *Science and Engineering Ethics*, 30(5), 47. <https://doi.org/10.1007/s11948-024-00513-0>

Link to the Creative Commons user license: <https://creativecommons.org/licenses/by/4.0/>

Conference presentations (of previous versions):

Poszler, F., Geisslinger, M. & Lütge, C. (May, 2023). A five-step ethical decision-making model for self-driving vehicles: Which (ethical) theories could guide the process and what values need further investigation? Presented at International Conference on Computer Ethics: Philosophical Enquiry 2023, Illinois Institute of Technology, Chicago. Available at: <https://www.youtube.com/watch?v=Un6ubQ2SYLA&list=PLQFndeOn8y10Xmj0eBXQYbQ4OZ9v8C0HW&index=2>

4.1 Introduction

Self-driving vehicles (SDVs) are one of the first commercialized AI-enabled robots to make decisions without human intervention. Although greater safety levels are attributed to SDVs than human drivers, they may end up in situations with fatal consequences for traffic participants (Rhim et al., 2020). While (ethical) decision-making of human drivers in traffic is rather intuitive, SDV's decision-making must be preprogrammed (Liu & Liu, 2021). Some opponents may argue that ethical computing (for dilemma situations) is irrelevant (UNECE, 2021); others point to its inevitability: "Even in instances in which no explicit ethical rule or preference is intended, the programming of a [SDV] may establish an implicit or inherent decision rule with significant ethical consequences" (U.S. Department of Transportation, 2016; p.26). This (implicit) programming has palpable effects on all road users in terms of risk impositions (Evans, 2021) or – in the worst case – traffic victims (Mordue et al., 2020). Therefore, SDV's decisions (i.e., trajectory selection), in any case, carry an ethical dimension and are of normative significance (Dietrich & Weisswange, 2019).

Considering the risks at hand and the fast-paced technological advancements, investigating the ethical programming of SDVs is a pressing concern (Nyholm & Smids, 2016), with the objective of potentially establishing a standardized, mandatory framework for such programming that serves the interests of society (Gogoll & Müller, 2017). Thus, many scholars have discussed what should feed into the programming and what constitutes ethical decision-making for SDVs. For example, to approach this conceptualization, Poszler et al. (2023) have conducted a holistic review of the autonomous driving ethics literature, in which they evaluated the applicability of certain ethical theories, identified additional considerations (such as situation-adjusted risk distributions) that may prove helpful to guide SDVs' ethical decision-making and synthesized ethical trajectory-planning algorithms that have been developed so far. Similarly, policymakers have recognized the importance of considering ethical dimensions in programming SDVs (e.g., European Commission, 2021; U.S. Department of Transportation, 2016). In this endeavor, compared to 'meaningless' and 'toothless' principles that should underlie AI development processes, such as fairness and privacy, operationalizing certain principles directly into product features seems to be a more promising approach (Munn, 2022). In the future, eliciting ethical values and turning them into concrete system requirements will be a regular task for system designers and so-called 'value leads' (compare IEEE 7000TM; IEEE, 2021). However, in terms of SDVs, "translating rules of the road, which are [amongst others] legal documents written in natural language, to formal rules for use by computers deployed on [autonomous vehicles] is a challenging task" (Bin-Nun et al., 2022; p.1).

Therefore, this article's objective is to support this endeavor by translating requirements recently formulated in policy drafts, standards, technical specifications, ethics guidelines, and scholarly work into an explicit five-step ethical decision model for SDVs during hazardous situations. We, hereby, aim to extend previous efforts by merging three key aspects: ethical, legal, and engineering considerations. For example, former scholarly work "rather focused on blending normative and empirical accounts of what matters morally in AV [i.e., autonomous vehicle] decision-making" (Evans, 2021; p.326) while transgressing restrictions that today have been legally specified, such as the discrimination of traffic participants based on personal features (e.g., age) (BMJ, 2023). Similarly, it has been suggested that engineering standards "may be reconsidered for future revision, including normative requirements" (SAE International, 2021a). With our simultaneous consideration of ethics, law, and engineering, we hope to achieve a higher level of social acceptance, ethical acceptability, legal substantiation, and technical feasibility of the resulting decision-making model for SDVs.

This paper is structured as follows. First, Section 4.2 highlights theoretical fundamentals drawn from policymakers, standardization organizations, and contemporary scholars. Second, Section 4.3 sketches the proposed model for the ethical decision-making of SDVs and elaborates its decision process with an exemplary traffic scenario. Third, the benefits and limitations of this model are illustrated, and underlying terms that need concretization in the future are pointed out (in Section 4.4). Lastly, a short conclusion is drawn. Overall, although not exhaustive and resolute, this article aims to serve the scholarly community by contributing to the debate on computational ethics and value-based engineering (especially in the field of autonomous driving) and practitioners in the automotive sector by laying out a potential solution for the 'ethical' programming of SDVs.

4.2 Theoretical fundamentals

This section provides a short overview of premises and requirements for an SDV's ethical decision-making that were raised by policymakers and scholars in the past and represents the fundament on which the subsequent proposed model is built. Overall, these requirements are fourfold: the management of risks (i.e., safety) (4.2.1), adjustments to the underlying calculation depending on the situation criticality (4.2.2), responsibility considerations and the protection of vulnerable road users (4.2.3) as well as implementing a mix of ethically grounded and socially shared principles (4.2.2).

4.2.1 Risk (i.e., safety) assessment and management

Generally, regulatory frameworks for AI systems point to assessing and managing pertinent applications using risk-based approaches (e.g., European Commission, 2021). The consideration of risks similarly exhibits centrality in regulatory drafts that specifically concern the functionality and programming of SDVs. In this context, risk is specified as the “combination of the probability of occurrence of harm and the severity of that harm” (ISO, 2018) or the product of collision probability and estimated harm (Geisslinger et al., 2021). Especially in the sight of and during critical traffic scenarios (defined in 4.2.2), SDVs are expected to calculate the probability and magnitude of imminent consequences (e.g., Dignum, 2019) and minimize risks to the safety of traffic participants by executing maneuvering to reach a ‘minimal risk condition’ (e.g., European Commission, 2022; Justia US law, 2022). This condition is characterized by ensuring the greatest possible road safety for all road users (e.g., ISO, 2023b; BMJ, 2023). Safety is understood in terms of physical integrity in that the protection of human life and reduction of road fatalities is given the highest priority over other considerations such as damage to property (e.g., BMJ, 2023; Kriebitz et al. 2022; Lütge et al., 2021). Furthermore, the safety/protection of passengers within the SDV must not be prioritized over third parties within the traffic scenario. Instead, even consideration of all road users is postulated (e.g., European Commission, 2022; Government of Canada, 2021; GOV.UK, 2023; NACTO, 2016). This commitment to foregoing risk inequality is further manifested by the demand for SDVs to “contribute to the reduction of the disproportional risk exhibited by certain road user groups” (Papadimitriou et al., 2022; p.3), which will be further elaborated in 4.2.3.

Key requirements for SDVs can be summarized as follows:

System requirements	Corresponding standards & regulations (Examples & further readings ²¹)
➤ Calculation of risks (i.e., product of collision probability and estimated harm)	• Definition of ‘risk’ in the context of SDVs (ISO, 2018)
➤ Performance of maneuvers to reach a ‘minimal risk condition’	• Constraints to avoid collisions with all other objects and to maintain a safe distance from other objects (GOV.UK, 2022) • Definition of ‘minimal risk condition’ and ‘minimal risk manoeuvre’ (DMV, 2022; European Commission, 2022; Government of Canada, 2021; ISO, 2021; ISO, 2022; Justia US law, 2022; Ministère Écologie

²¹ Exemplary references to standards and regulations that address the ‘ethical’ decision-making of SDVs, while general regulations or standards that did not concern SDVs’ programming were excluded, such as the U.S. Department of Transportation (2021). All standards and regulations consulted in this article are listed in Appendix of Essay III – List of relevant standards & regulatory documents. This list should not be understood as an exhaustive compilation but as a guide outlining some key aspects of existing law and established standards.

➤ Minimal risk=greatest possible safety (i.e., physical integrity of human beings)	Énergie Territoires, 2022; SAE International, 2021b; U.S. Department of Transportation, 2016)
➤ Simultaneous consideration of the safety of traffic participants in and outside the SDVs	<ul style="list-style-type: none"> • Definition of ‘minimal risk’ (ISO, 2023b; BMJ, 2023) • Definition of ‘safety’ in the context of SDVs (i.e., in terms of physical integrity and road fatalities) (BMJ, 2023) • Consideration of all road users beyond passengers (i.e., third parties, the broader transport ecosystem) (BMJ, 2023; European Commission, 2022; Government of Canada, 2021; GOV.UK, 2023; NACTO, 2016; UNECE, 2022)

4.2.2 Adjustments to the underlying calculation depending on the situation's criticality

In general, SDVs are expected to increase road safety (Lütge, 2017) and, thus, mostly encounter collision-free situations. Nevertheless, not all risks can be bypassed by the introduction and operation of SDVs (Goodall, 2016), so collisions may emerge. Correspondingly, this duality of traffic scenarios has been recognized by introducing terms such as ‘non-hazard’ and ‘hazard’ situations (Dietrich & Weisswange, 2019) or ‘non-critical’ and ‘critical’ occurrences (UNECE, 2022). In contrast to ‘non-critical’ situations, ‘critical situations’ entail that “at least one person suffers an injury that requires medical attention” (UNECE, 2022; p.32; Ministère Écologie Énergie Territoires, 2022). The distinction between these types can be based on certain metrics and risk thresholds and be followed by a remedial action in case an unacceptable risk is predicted (UNECE, 2022). Such an ad hoc action in ‘critical’ situations is characterized by selecting the trajectory that protects human life above everything else, such as damages to animals or property (Zhu, 2021) or strict compliance with road traffic laws (e.g., crossing a solid line to pass a cyclist) (NTC, 2022). Therefore, human life (i.e., physical integrity) is the only factor entering the outcome calculation when comparing different trajectories. On the other hand, in ‘non-critical’ situations, additional factors can be consulted when determining the optimal trajectory, such as mobility or passenger comfort (e.g., Dietrich & Weisswange, 2019; Geisslinger et al., 2023a; Westhofen et al., 2023). Thus, the categorization of traffic situations is decisive for the data consulted in the SDV’s calculation and decision-making process.

Key system requirements can be summarized as follows:

System requirements	Corresponding standards & regulations
---------------------	---------------------------------------

(Examples & further readings)	
➤ Separation of traffic situations into ‘hazard’ and ‘non-hazard’ and corresponding SDV responses	<ul style="list-style-type: none"> • Distinction of situations into ‘non-critical’ and ‘critical’ occurrences or approaches (ISO, 2023a; UNECE, 2022) or ‘normal driving’ and ‘crash avoidance’ responses (U.S. Department of Transportation, 2016) • Definition of ‘dilemmas’ as critical situations (European Commission, 2020) • Definition of ‘critical situations’ and ‘personal road traffic accidents’ (European Commission, 2022; UNECE, 2022; Ministères Écologie Énergie Territoires, 2022) • Definition of ‘hazardous situation’ (IEEE, 2022; ISO, 2020a) • Definition of ‘conflict zone’ (ETSI, 2018)
➤ Criticality determination based on metrics, thresholds relating to unacceptable risk	<ul style="list-style-type: none"> • Emphasis of a minimum threshold level of safety and unacceptable risk (Government of Canada, 2021; UNECE, 2022) • Unacceptable risk depends on factors such as the level of controllability from other road users, rates of occurrence, or severity levels of the outcome of a particular traffic scenario (IEEE, 2022) • Predefined collision risk probability determines the need to initiate actions to avoid a collision (ETSI, 2018; ETSI, 2021b) • Indicators for the criticality of the traffic safety situation, such as Time-to-collision or the time required to act to avoid or mitigate a collision (ETSI, 2018; ETSI, 2021b; ISO, 2023a)
➤ Consideration of a single parameter (i.e., human life) in ‘hazard’ situations	<ul style="list-style-type: none"> • Protection of human life as the highest priority compared to other legal interests (BMJ, 2023) • Primary road safety application is to prevent collisions (ETSI, 2018; ETSI, 2021b) • Examples of obligations that can be neglected, e.g., strict compliance with road traffic law (NTC, 2022)

4.2.3 Responsibility considerations and the protection of vulnerable road users

As stated in 4.2.1., risk distributions are, in principle, to be allocated equally between all traffic participants. Berkey (2022) endorses this with a side constraint, namely, “unless there is a morally compelling reason for deviating from this aim” (p.11), such as allocating a greater share of risks to those road users who introduce risks in the first place. This aligns with the concept of moral responsibility, which holds that drivers, despite any precautions, are responsible for causing harm in the event of a collision since they voluntarily engage in activities that threaten others (Kauppinen, 2021). By contrast, cyclists or pedestrians impose much less risk on road traffic due to their lower mass and velocities (Geisslinger et al., 2023a). Therefore, although owners of SDVs

may not be liable for causing an accident (Jensen, 2018), SDVs could be programmed to assume higher levels of risk compared to other road users in critical situations. In practice, a vulnerability categorization of the road users is proposed, in which cyclists, pedestrians, or generally, individuals outside the vehicle are distinguished from individuals inside the SDV (Evans et al., 2020). The necessity to protect vulnerable road users and, hence, balance risks between different classes of road users is similarly acknowledged by regulatory bodies (e.g., European Commission, 2022; Government of Canada, 2021; GOV.UK, 2022/2023; NHTSA, 2022).

Key requirements can be summarized as follows:

System requirements	Corresponding standards & regulations (Examples & further readings)
➤ In principle, equal treatment of all traffic participants	• Equality in the safety of all road users (European Commission, 2020)
➤ Special protection for vulnerable traffic participants	• Commensurate level of safety for vulnerable road users (European Commission, 2020; GOV.UK, 2022; ISO, 2023b; NHTSA, 2022)
➤ Categorization of road users in line with their relative vulnerability	• Definition of ‘vulnerable’ and ‘non-vulnerable’ road users (ISO, 2020b) • Definition and characterization of ‘vulnerable road users’ in terms of parameters such as speed and weight class (ETSI, 2021b)
➤ Potential categorization into vehicles, cyclists, pedestrians, or users inside and outside an SDV	• List/examples of ‘vulnerable road users’ (ETSI, 2018/2021; European Commission, 2020/2022; Government of Canada, 2021; IEEE, 2022; NHTSA, 2022)

4.2.4 Implementing a mix of ethically grounded and socially shared principles

To achieve ethical decision-making and behavior of autonomous systems, it is important to integrate society’s values and ethical principles within the underlying algorithms (Dyoub et al., 2020). Ethical principles are here understood as “[o]perationalizable rules inferred from philosophical theories such as Deontology and Consequentialism” that help determine the moral permissibility of an action (Woodgate & Ajmeri, 2022; p.3). Next to these normative theories, societal preferences can inform the decision-making logic of SDVs (Poszler et al., 2023). Specifically for SDVs, scholars highlight the need to create ‘mixed’ algorithms that combine various normative ethical theories (Geisslinger et al., 2021; Hübner & White, 2018) and society’s values (Robinson et al., 2021). For example, Evans et al. (2023) propose using Kantian, Millian, or descriptive ethics as sources of inspiration for restrictions and mitigation rules that can be integrated into the ‘soul’ (i.e., the algorithm) of SDVs. This allows accounting “for a variety of ethical concerns [a]nd [...] achieve widespread acceptance in society” (Evans, 2021; p.324).

A mix of normative ethical theories is proposed since unfair outcomes of relying on a single theory may be mitigated by utilizing a ‘balanced approach’ (ISO, 2023b) or ‘pluralistic approach’, in which “a variety of principles can be weighed against one another in order to find the fairest answer” (Woodgate & Ajmeri, 2022; p.12). For example, there may be instances where it becomes unreasonable to follow strict duties, such as prioritizing the absolute protection of bystanders such as pedestrians (as indicated as a requirement in 4.2.3). Such a notion of reasonableness could allow accepting minimal chances of harming an individual if, as a result, another person is saved with certainty (Sütfeld et al., 2019). Similarly, it might be justifiable to deviate from aiming at an equal distribution of the risks (as indicated as a requirement in 4.2.3) to reduce the absolute level of risk (i.e., harm) that every traffic participant is subjected to (Berkey, 2022). Naturally, conflicts and inconsistencies will emerge when consulting various normative theories simultaneously. For example, in the previously mentioned case, the level of risk imposition to one individual may compete with the aggregated level of risk for all traffic participants. When such conflicts emerge, humans can resolve these situations by accepting tradeoffs, developing hierarchical relations, or assigning weights (IEEE, 2019). Therefore, regulatory bodies emphasize managing the decision-making of SDVs by “shared ethical principles” that align with societal values and preferences (European Commission, 2020; p.7). However, preferences indicated in empirical studies cannot be blindly adopted due to “the risk of committing the naturalistic fallacy” (Jacobs & Hultgren, 2021; p.24). For example, in the Moral Machine experiment, participants stated they preferred sacrificing people who are old, overweight, or homeless in accident situations involving SDVs (Awad et al., 2018). Even if this data could be estimated using sensor measurements (Németh, 2023), the European Commission (2020) prohibits the discrimination of humans based on their personal characteristics in critical situations. Thus, a promising approach seems to complement normative theories with societal values that comply with regulations (Dignum, 2019).

Key requirements can be summarized as follows:

System requirements	Corresponding standards & regulations
<ul style="list-style-type: none"> ➤ Integration of a mix of normative theories, such as deontological and consequentialist ethics ➤ Consideration of the reasonableness of risk impositions 	<p>(Examples & further readings)</p> <ul style="list-style-type: none"> • Balanced approach that emphasizes the consideration of different normative ethical theories (ISO, 2023b) • Freedom of unreasonable safety risks (European Commission, 2022; ISO, 2022; NHTSA, 2022; UNECE, 2022; U.S. Department of Transportation, 2016) • Safety = absence of unreasonable risk (ISO, 2020b; ISO, 2018)

➤ Alignment with society's values via manifesting their preferred hierarchical orders, thresholds, or weights	<ul style="list-style-type: none"> • Importance of 'shared' ethical principles when managing risk distributions (European Commission, 2020) • Acceptance criterion that derives unreasonable levels of risk from 'valid societal moral concepts' (ISO, 2022)
➤ Prohibition to discriminate based on personal characteristics of humans	<ul style="list-style-type: none"> • Risk distributions based on personal characteristics are forbidden (BMJ, 2023) • Risk calculations should instead be based on physical properties such as the dynamic state and mass of the objects (ETSI, 2021b)

4.3 A proposed model for ethical decision-making of SDVs

Based on the groundwork stated in the previous section, this paper proposes a five-step model for ethical decision-making of SDVs and elaborates its decision process with an exemplary, simplified traffic scenario. The overall sequence of steps is illustrated in Figure 25. All relevant terms and examples of its technical measures/indicators are defined in Table 17.

Step 1: Determination & calculation of possible trajectories. Decisions of SDVs are implemented via trajectory planning and selection. Thus, in the first step, the SDV needs to determine all potential trajectories and calculate corresponding consequences for each trajectory alternative (A_i) and each traffic participant (T_i). Consequences that play a role in road traffic include, first and foremost, safety (i.e., the physical integrity of the traffic participants, determined by the risk posed to them). The risk for each traffic participant ($r_{Ai,Ti}$) can be defined as the product of collision probability ($c_{Ai,Ti}$) and estimated harm ($h_{Ai,Ti}$). Additional – yet subordinated – utilities or objectives ($x_{Ai,Ti}$) of SDVs include passengers' comfort or mobility. The calculation that is necessary for this step is grasped by Equation 1²²:

$$r_{Ai,Ti} = c_{Ai,Ti} h_{Ai,Ti} \quad (1)$$

Exemplary elaboration. In the imagined traffic scenario (Figure 24), the SDV (T_1) has four trajectory alternatives, which are: (A_1) collide with an oncoming vehicle (T_2) to the left, (A_2) collide with a vehicle (T_3) in front, (A_3) collide with a pedestrian (T_4) on the sidewalk or (A_4) crash into a wall²³.

²² In this paper, we will not provide an equation for calculating utilities/goals other than safety and take the numerical figure for $x_{Ai,Ti}$ in Table 9 as given. This is because this paper focuses on the 'hazard mode', where the key objective of SDVs is safety. Please see Table 17 for additional resources on calculating utilities such as mobility or comfort.

²³ While the imagined traffic scenario assumes the existence of T_2 , T_3 , T_4 , and the wall, these individuals/objects are not understood as certainties. This is accounted for by factoring in the collision probability, which results from uncertainties in autonomous driving, such as the actual localization of traffic participants (Geisslinger et al., 2023a).

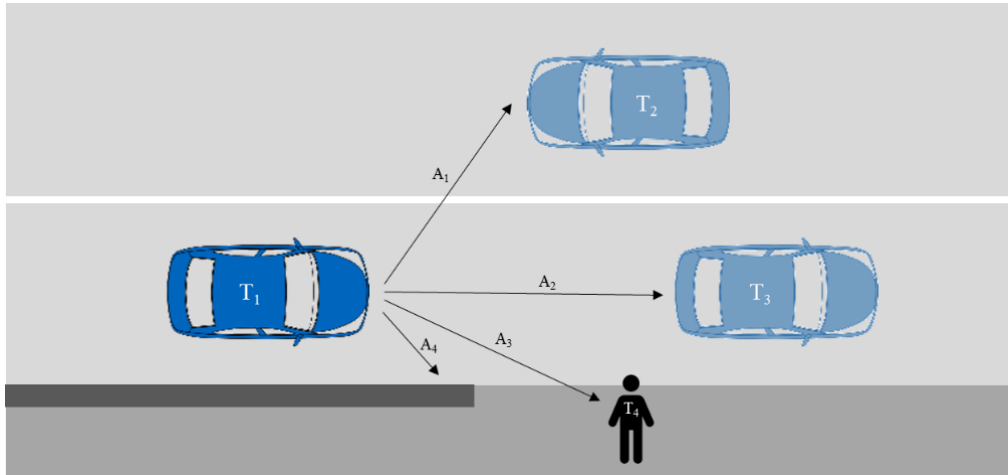


Figure 24: Simplified traffic scenario

For each of the four trajectories, the SDV calculates consequences, for example, the risk for each traffic participant ($r_{Ai,Ti}$) and any additional relevant utilities/objectives ($x_{Ai,Ti}$), as illustrated in Table 9. In this example, the numerical figures for risk range from 0 (i.e., no collision, entailing no injury) to 1 (i.e., certain collision, entailing a fatal injury). Similarly, the numerical figures for $x_{Ai,Ti}$ range from 0 (e.g., no comfort) to 1 (e.g., complete comfort).

Trajectory alternatives (A_i)	Traffic participants (T_i)									
	T_1			T_2		T_3		T_4		
	$c_{Ai,T1}$	$h_{Ai,T1}$	$x_{Ai,T1}$	$c_{Ai,T2}$	$h_{Ai,T2}$	$c_{Ai,T3}$	$h_{Ai,T3}$	$c_{Ai,T4}$	$h_{Ai,T4}$	
	$r_{Ai,T1}$			$r_{Ai,T2}$		$r_{Ai,T3}$		$r_{Ai,T4}$		
A_1	0.50	0.50	0.5	0.60	0.40	0.00	0.00	0.00	0.00	
	0.25			0.24		0.00		0.00		
A_2	0.50	0.70	0.9	0.10	0.20	0.50	0.50	0.00	0.00	
	0.35			0.02		0.25		0.00		
A_3	0.90	0.20	0.8	0.00	0.00	0.00	0.00	0.80	0.90	
	0.18			0.00		0.00		0.72		
A_4	0.90	0.60	0.4	0.00	0.00	0.00	0.00	0.10	0.40	
	0.54			0.00		0.00		0.04		

Table 9: Overview of risks ($r_{Ai,Ti}$) and an unspecified utility/objective ($x_{Ai,Ti}$) for all trajectory alternatives²⁴ and the individual traffic participants

²⁴ Note: Strictly speaking, the collision probabilities for the traffic participants that are potentially colliding with each other are symmetrical. To elaborate, for trajectory A_3 , where the SDV (T_1) potentially collides with the pedestrian (T_4), their collision probabilities should be equal. However, Table 9 presents varying collision probabilities for these two

Step 2: Typification of situation. The SDV determines the nature of the situation based on its ability to fulfill particular key duties. As the prime requirement for SDVs is safety, key duties (to be prioritized over utilities or objectives such as comfort or mobility) entail safeguarding the physical integrity of all traffic participants. Exemplary rules/duties are displayed in Table 10, which will be consulted when assessing traffic situations and consequences such as those illustrated in Table 9. If the SDV concludes that at least one of the established rules/duties will be disobeyed (determined by the surpassing of a particular risk figure), the SDV will declare a ‘hazard situation mode’ implicating a specific decision-making process (that is different from the ‘non-hazard situation mode’²⁵). Namely, in the ‘hazard situation mode’, the only consequence to be contemplated is risk, while other consequences, such as the passenger’s comfort or mobility, are to be neglected.

If any of the following **duties** cannot be fulfilled, the SDV declares a ‘**hazard situation**’ mode:

1. The lives of other traffic participants must not be put in harm’s way. (e.g., $r_{Ai,T2}$ must not > 0.2)²⁶
 2. The lives of SDV passengers must not be put in harm’s way. (e.g., $r_{Ai,T1}$ must not > 0.2)
-

Table 10: Exemplary SDV duties distinguishing between non-hazard and hazard modes, adapted from Evans et al. (2020)

Exemplary elaboration. In the imagined situation (Figure 24), all four trajectory alternatives entail risk above the level of 0.2 for each traffic participant (as illustrated in Table 9). According to the commands in Table 10, this means the SDV cannot fulfill its duty to not cause harm to traffic participants and/or the passenger(s) of the SDV. Therefore, the SDV will switch to ‘hazard situation mode’, implicating that utilities/objectives such as comfort or mobility are no longer directive for the subsequent calculations.

colliding parties (i.e., $c_{Ai,T1}=0.9$ and $c_{Ai,T4}=0.8$). This discrepancy is due to mutual influences among all traffic participants and other existing obstacles that we assume to know in this article.

²⁵The further course of action for the ‘non-hazard mode’ is not further elaborated here since the focus of this article is SDVs’ decision-making in hazardous situations. Potential steps in the ‘non-hazard mode’ include balancing all consequences, such as safety, comfort, and mobility.

²⁶ The indicated numerical figures throughout this section are not to be taken at face value; they are only stated as examples for explanatory purposes. In reality, to compete with human driving performance, risks must be much smaller, e.g., the EU legislation uses a value of 10^{-7} fatalities per hour as an example. The discussion section of this article highlights how to approximate the actual numerical figures that are to be utilized as reference points in practice.

Step 3: Exclusion of prohibited trajectories. To identify prohibited trajectories, the SDV checks the consequences of all trajectory alternatives for every traffic participant against particular risk thresholds. Thresholds to be contemplated are those for collision probability and estimated harm, each separately (i.e., c_{max} and h_{max}). As indicated in Table 11: Exemplary threshold restrictions for collision probability (c_{max}) and estimated harm (h_{max})

, if the collision probability exceeds a particular numerical figure for one traffic participant, this individual's estimated harm must not exceed a particular threshold. A similar logic applies if the estimated harm for a traffic participant exceeds a certain numerical figure. Those trajectories that fail to fulfill particular threshold restrictions are to be excluded; all remaining trajectory alternatives are reevaluated by the SDV's algorithm in step 4.

If $c_{Ai,Ti} > 0.8$, $h_{Ai,Ti}$ must not $> h_{max}=0.7$
OR
If $h_{Ai,Ti} > 0.8$, $c_{Ai,Ti}$ must not $> c_{max}=0.6$

Table 11: Exemplary threshold restrictions for collision probability (c_{max}) and estimated harm (h_{max})

Exemplary elaboration. According to the calculated consequences in Table 9 (i.e., $c_{A3,T4}=0.80$ and $h_{A3,T4}=0.90$) and the threshold restrictions in Table 11, the SDV has to exclude the trajectory alternative A_3 since for traffic participant T_4 none of the two restrictions is adhered to. Therefore, the SDV's new action space for further consideration is limited to trajectory alternatives A_1 , A_2 , and A_4 .

Step 4: Calculation of valence-adjusted risk. The SDV reevaluates all remaining trajectory alternatives by adjusting the risk figures with valence factors (v_i) for the different traffic participants. This valence factor corresponds to the traffic participant's vulnerability. For example, traffic participants could be classified into pedestrians, cyclists, and vehicles, with gradually declining valence factors, as illustrated in Table 12. The calculation that is necessary for this step is grasped by Equation 2:

$$vr_{Ai,Ti} = v_i r_{Ai,Ti} \quad (2)$$

The new valence-adjusted risk figures (i.e., $vr_{Ai,Ti}$) (illustrated in Table 13) feed into the decision-making process of Step 5.

For pedestrians, $v_{ped}=1.0$

For cyclists, $v_{cyc}=0.8$

For motor vehicles,

$v_{mot}=0.5$

Table 12: Exemplary hierarchy and corresponding valence factors (v_i) for different types of traffic participants

Exemplary elaboration. In the imagined, simplified scenario (Figure 24), there are only two types of traffic participants, i.e., one pedestrian and motor vehicles. Therefore, utilizing the predetermined valence factors in Table 12 (i.e., v_{ped} and v_{mot}), the SDV calculates the valence-adjusted risk ($vr_{Ai,Ti}$) for each traffic participant in each of the three remaining trajectory alternatives A_1 , A_2 and A_4 . This implies that the new risk figures for the involved vehicles (i.e., $vr_{Ai,T1}$, $vr_{Ai,T2}$, $vr_{Ai,T3}$) are lower than their initially calculated risk (i.e., $r_{Ai,T1}$, $r_{Ai,T2}$, $r_{Ai,T3}$). In contrast, the new risk figures for the involved pedestrian (i.e., $vr_{Ai,T4}$) are not discounted but remain the same due to the valence factor for pedestrians of 1.0.

Trajectory alternatives (A_i)	Traffic participants											
	(T_i)											
	T_1			T_2			T_3			T_4		
	$r_{Ai,T1}$	v_{mot}	$vr_{Ai,T1}$	$r_{Ai,T2}$	v_{mot}	$vr_{Ai,T2}$	$r_{Ai,T3}$	v_{mot}	$vr_{Ai,T3}$	$r_{Ai,T4}$	v_{ped}	$vr_{Ai,T4}$
A_1	0.25	0.50	0.13	0.24	0.50	0.12	0.00	0.50	0.00	0.00	1.00	0.00
A_2	0.35	0.50	0.18	0.02	0.50	0.01	0.25	0.50	0.13	0.00	1.00	0.00
A_4	0.54	0.50	0.27	0.00	0.50	0.00	0.00	0.50	0.00	0.04	1.00	0.04

Table 13: Overview of calculated valence-adjusted risks ($vr_{Ai,Ti}$) for all remaining trajectory alternatives and the individual traffic participants

Step 5: Selection of final trajectory. This step aims to identify the one trajectory that meets two risk distribution principles, namely, the greatest equal risk between traffic participants and that optimizes (i.e., minimizes) aggregated risk. Thus, based on the valence-adjusted risks, the SDV calculates the risk inequality (E_{Ai}) between all traffic participants and the aggregated risk (U_{Ai}) for all trajectory alternatives. The first calculations that are necessary here are grasped by Equations 3 and 4:

$$E_{Ai} = \sum |vr_{Ai,Ti} - vr_{Ai,Tj}| \quad (3)^{27}$$

$$U_{Ai} = \sum vr_{Ai,Ti} \quad (4)$$

The degree to which the risk distribution principles E_{Ai} and U_{Ai} are factored in and are decisive for the selection of the final trajectory is predetermined with a weighting factor for each principle (i.e., w_E and w_U). Exemplary weighting factors are provided in Table 14.

$$w_E=0.5$$

$$w_U=0.5$$

Table 14: Exemplary weighting factors for risk inequality (w_E) and aggregated risk (w_U)

Given these weightings, the SDV can select the final action, i.e., the trajectory with the lowest principle-weighted risk (wr_{Ai}). The corresponding calculation is grasped by Equation 5:

$$wr_{Ai} = w_E E_{Ai} + w_U U_{Ai} \quad (5)$$

Exemplary elaboration. Table 15 illustrates the calculated risk inequality and aggregated risk for each remaining trajectory alternative. In the imagined situation, the first trajectory alternative (A_1) satisfies the utilitarian principle most because it offers the lowest aggregated risk ($E_{A1}=0.25$). Furthermore, this trajectory alternative performs best concerning the Equality principle as it entails the lowest risk inequality ($U_{A1}=0.51$) between all parties involved.

Trajectory alternatives (A_i)	Traffic participants (T_i)				Risk distribution principles	
	T_1	T_2	T_3	T_4	E_{Ai}	U_{Ai}
	$vr_{Ai,T1}$	$vr_{Ai,T2}$	$vr_{Ai,T3}$	$vr_{Ai,T4}$		
A_1	0.13	0.12	0.00	0.00	0.51	0.25
A_2	0.18	0.01	0.13	0.00	0.66	0.32
A_4	0.27	0.00	0.00	0.04	0.85	0.31

Table 15: Overview of calculated risk inequality (E_{Ai}) and aggregated risk (U_{Ai}) for all remaining trajectory alternatives

²⁷ The risk disparities across all combinations of traffic participants will be summed up to determine the risk inequality for a particular trajectory alternative.

As a next step, the SDV has to calculate the principle-weighted risk figures by drawing on the predetermined weighting factors (as illustrated in Table 14). As an example, the weighting factors for both – risk inequality (w_E) and the aggregated risk (w_U) – are here both 0.5²⁸. Therefore, as illustrated in Table 16, the SDV will conclude that A_1 will entail the best outcome since this trajectory results in the lowest principle-weighted risk figure ($wr_{A1}=0.38$). Ultimately, the SDV will select A_1 for action execution.

Trajectory alternatives (A_i)	Weighted risk distribution principles		Principle-weighted risks (wr_{Ai})
	$w_E E_{Ai}$	$w_U U_{Ai}$	
A_1	0.255	0.125	0.38
A_2	0.33	0.16	0.49
A_4	0.425	0.155	0.58

Table 16: Overview of calculated principle-weighted risks (wr_{Ai}) for all remaining trajectory alternatives

²⁸ If other weighting factors were utilized here, the preferred decision may diverge between Table 15 and Table 16.

Terms	Definitions	Technical measures/indicators (Examples & further readings ²⁹)
Trajectory alternatives (A)	Set of possible actions/trajectories (physically) available to an SDV (e.g., Evans, 2021; Geisslinger et al., 2023b)	<ul style="list-style-type: none"> • Driveable area (Lin & Althoff, 2023) • Potential fields for crossable and non-crossable obstacles and road boundaries (Wang et al., 2020) • Reachable sets (Coskun, 2021)
(A _i)	Trajectory <i>i</i>	
Traffic participants (T)	Set of all traffic participants in a scenario	
(T _i)	Traffic participant <i>i</i>	
Risk (r_{A_i, T_i})	Combination of the probability of occurrence of harm and the severity of that harm (e.g., ISO, 2018)	<ul style="list-style-type: none"> • Product of collision probability and estimated harm (D'Souza et al., 2023; Geisslinger et al., 2021; Geisslinger et al., 2023a/b; Trauth et al., 2023) • Pedestrian risk index (Westhofen et al., 2022)
Collision probability (c_{A_i, T_i})	Likelihood of a collision happening	<ul style="list-style-type: none"> • Lateral and longitudinal separation distance between two vehicles (D'Souza et al., 2023) • Time-to-reach (Aksjonov & Kyrki, 2023); Time-to-collision; Time-to-react (Abdelhalim & Abbas, 2022; Geisslinger et al., 2023b; Lin & Althoff, 2023; Westhofen et al., 2022; Wishart et al., 2020) • Crash potential index; aggregated crash index (Lin & Althoff, 2023; Westhofen et al., 2022) • Proximity to the predicted trajectory of other traffic participants (Evans, 2021)
Estimated harm (h_{A_i, T_i})	Severity of a collision in terms of the damage to the physical integrity of a human being (e.g., Evans, 2021; Geisslinger et al., 2021)	<ul style="list-style-type: none"> • Delta-v – change in velocities following the collision (Evans, 2021; Evans et al., 2023; Lin & Althoff, 2023; Robinson et al., 2021) • Velocities and masses of colliding traffic participants (Evans, 2021; Geisslinger et al., 2023a; Robinson

²⁹ To pinpoint relevant readings, the structured literature review by Poszler et al. (2023) served as a starting point. Specifically, we have included the ten publications identified in this review, which discuss 'elaborate decision processes' (e.g., ethical trajectory-planning algorithms) for SDVs, along with supplementary readings that cited these ten publications.

Essay III: Ethical Decision-Making for Self-Driving Vehicles: A Proposed Model & List of Value-Laden Terms That Warrant (Technical) Specification

		<p>et al., 2021; Trauth et al., 2023; Westhofen et al., 2022)</p> <ul style="list-style-type: none"> • Impact areas and/or angles (Evans, 2021; Geisslinger et al., 2023a; Trauth et al., 2023) • Potential crash severity index (PCSI) determined by factors such as characteristics of the obstacles, approaching velocity, relative crash angle, and weight difference (Wang et al., 2020) • Collision harm index determined by velocities and masses of colliding traffic participants as well as these corresponding SDV properties post-collision (D'Souza et al., 2023) • Risk of fatality; Maximum Abbreviated Injury Score (MAIS) (Evans et al., 2023; Geisslinger et al., 2023a; Robinson et al., 2021; Trauth et al., 2023) • KABCO injury classification scale (Wishart et al., 2020)
Additional utilities/objectives ($x_{Ai,Ti}$)	Outcome figures for SDVs' desirable goals, next to safety	<ul style="list-style-type: none"> • Comfort or mobility (determined by, for example, the vehicle's acceleration and jerk) (Geisslinger et al., 2023a/b) • Efficiency (determined by velocity), comfort (determined by longitudinal and lateral acceleration) (Aksjonov & Kyrki, 2023) • Cost function that integrates objectives such as comfort or energy consumption (Németh, 2023) • Path tracking or occupant comfort (determined by change in steering of lateral front tire force) (Thornton et al., 2016)
Maximum acceptable collision probability (c_{max})	Threshold for collision probability that must not be exceeded, given a particular estimated harm figure is exceeded	
Maximum acceptable estimated harm (h_{max})	Threshold for estimated harm that must not be exceeded, given a particular collision probability figure is exceeded	<ul style="list-style-type: none"> • Current safety levels with human drivers in conventional cars (Geisslinger et al., 2023a) • MAIS3+ level injury (Evans et al., 2023)

Traffic participant valence (v_i)	Weighting factor for traffic participant i according to their relative vulnerability	
Valence-adjusted risk ($vr_{Ai,Ti}$)	Risk figure adjusted by the particular valence of traffic participant i	
Risk inequality (E_{Ai})	Sum of risk differences among all traffic participants in trajectory i	<ul style="list-style-type: none"> Equality principle (Geisslinger et al., 2021; Geisslinger et al., 2023a); egalitarian approach (Evans et al., 2023)
Aggregated risk (U_{Ai})	Sum of expected harms for all traffic participants in trajectory i (e.g., Evans et al., 2023)	<ul style="list-style-type: none"> Bayesian principle (Geisslinger et al., 2021; Geisslinger et al., 2023a); utilitarian approach (Evans et al., 2023)
Risk inequality weighting (w_E)	Weighting factor for the distribution principle relating to risk inequality	
Aggregated risk weighting (w_U)	Weighting factor for the distribution principle relating to aggregated risk	
Principle-weighted risk (wr_{Ai})	Risk figure that balances principles related to risk inequality and aggregated risk according to their particular weighting factors	<ul style="list-style-type: none"> Risk-cost function that balances different ethical principles (Geisslinger et al., 2021; Geisslinger et al., 2023a)

Table 17: Definition and technical specification of key terms for the SDVs' decision steps

4.4 Discussion

In this section, we will highlight the benefits of the previously sketched decision-making process (4.4.1) and its limits and point to future research by listing (how to determine) the underlying terms that need concretization (4.4.2).

4.4.1 Benefits of this ethical decision-making model

Having in mind the previously sketched requirements for an SDV's ethical decision-making process (see 4.2), our proposed model works towards generating the following benefits:

- ✓ **The proposed model provides a chronological order, consistency, and justifications for particular decision-making steps while leaving room for adjustments.** This model states a precise sequence of steps an SDV could engage in and execute an ethical decision-making process (i.e., select the most appropriate trajectory). The whole decision procedure is elaborated in more detail by running the steps through an imagined traffic scenario. Since the model is derived from the theoretical fundamentals stated in Section 4.2, its basic structure aims to – as proposed by Dignum (2019) – align with existing regulations, standards, normative ethical theories, and societal values and, thereby, derive explanatory and justificatory power (Jacobs & Hultgren, 2021). While its basic structure of five steps is here assumed as static, the model offers many possibilities for adjustments by, for example, not putting the list or figures for duties, thresholds, or weightings (e.g., c_{max} , w_E) in concrete numerical terms.
- ✓ **The proposed model utilizes overall risk (i.e., safety) as a key factor (starting with step 1).** In line with reality, in which the occurrence of events during traffic operations is uncertain (Liu, 2017) and in line with previously stated policymakers' calls, this model adopts risks (i.e., the product of calculation of collision probability and estimated harm) as a key pillar from the beginning. As suggested in Section 4.2, this model showcases an exemplary process for an SDV maneuver to reach a 'minimal risk condition', which concentrates on safety as an ultimate aim. Furthermore, the underlying risk distribution considers not only the SDV's passengers (e.g., as illustrated in Figure 24, the SDV, two other vehicles, and a pedestrian are considered), thereby striving for equal consideration of all involved road users as a starting point.
- ✓ **The proposed model adapts to the context at hand (e.g., in step 2).** Based on the SDV's ability to fulfill certain duties that are linked to particular risk metrics, traffic situations are separated into a 'hazard situation mode' or 'non-hazard situation mode'. This typification is decisive for the following decision-making process. For example, as suggested in Section 4.2.2, once a traffic situation is identified as critical, the proposed model relies on the single

parameter of safety (i.e., physical integrity) in its subsequent calculations. Other parameters, such as passenger comfort or mobility, could be accounted for in less critical traffic situations. Therefore, this model responds to situational demands and context information to choose between ethical principles (Woodgate & Ajmeri, 2022).

- ✓ **The proposed model accounts for the reasonableness of risk impositions (e.g., in steps 3 and 5).** As stated in Section 4.2.2, it may not be reasonable to forego trajectories that could entail fatal harm at any price, especially when the occurrence of a collision is unlikely in the first place. Therefore, in step 3, this model aims to manifest such reasonableness of risk-taking by instantiating a ‘conditional Maximin strategy’, which means that, for example, a high level of estimated harm is only accepted if a low collision probability accompanies it. Therefore, trajectories that entail high levels of estimated harm are not per se to be rejected. The notion of reasonableness is further addressed in step 5, in which different ethical theories (such as the Equality principle and utilitarianism) are weighted to – in line with suggestions made by Berkey (2022) – allow the consideration and reduction of aggregate risk within a traffic scenario.
- ✓ **The proposed model incorporates considerations of responsibility and the protection of vulnerable road users (e.g., in step 4).** In step 4 of this model, traffic participants are categorized into road user types in line with their relative vulnerability. Similar to existing propositions (compare 4.2.3), road user types include vehicles, cyclists, and pedestrians. For these three types, corresponding valence factors are allocated, whereby pedestrians receive the highest and vehicles receive the lowest valence factor. Therefore, this step ensures special and double³⁰ protection of vulnerable traffic participants and the incorporation of responsibility because the SDV algorithm acknowledges and compensates for the varying levels of risk that different parties introduce to traffic in the first place. Namely, the selected trajectory (A_1) warrants a high distance between the SDV and the existing vulnerable traffic participant (T_4).
- ✓ **The proposed model relies on a mix of ethically grounded and socially shared principles (e.g., in step 5).** Although this approach can be considered structurally consequentialist (Evans, 2021), a plurality of (ethical) theories make up the body of the decision-making process, namely, risk management in the form of outcome calculations and thresholds, deontological ethics in the form of duties, the Maximin principle in the form of a constraint, utilitarianism (i.e., by minimizing aggregated risk) or the Equality principle (i.e., by reducing

³⁰ Note: ‘Double’ protection for vulnerable road users results because, by nature, their inherent risk figures are already elevated due to their higher estimated harm. In step 4, these risk figures will be further heightened compared to those of other road users, serving as an additional buffer.

risk inequality) in the form of distribution strategies that are weighted against each other. In addition, with the many figures (e.g., c_{max} , v_i , w_E) being left open to be fixated, this model leaves space for the alignment with societal values and preferences (e.g., about unacceptable risk thresholds, desired hierarchical order of road user types or weights for particular risk distribution strategies). By generally providing opportunities to combine ethical theories with insights from descriptive ethics, this model aims to forgo “the risk of attending to a set of values that is unprincipled or unbounded” (Jacobs & Hultgren, 2021; p.23).

4.4.2 Limitations, research agenda, and terms to be determined

Despite the previously sketched benefits, this proposed decision-making model has limitations. The following issues warrant careful consideration and demand further investigation:

1. **Potential technical issues, inherent biases, and tradeoffs.** From a technical perspective, this model is an abstraction from reality. For example, our decision-making model assumes SDVs have the capability to identify hazardous situations and act appropriately in a prompt manner. However, ethical computation and motion planning processes might require more time (e.g., 2 ms per trajectory) than state-of-the-art algorithms that do not integrate ethical considerations (Geisslinger et al., 2023). This may limit the SDVs’ ability to respond in real time. Therefore, future research should investigate whether the extra time required to undergo the suggested decision-making steps is practically achievable in due time without risking a collision. Moreover, although this model explicitly refrains from discriminating traffic participants based on personal characteristics and seeks to protect vulnerable road users, inherent biases may nevertheless creep in and influence the SDV’s outcomes. For example, AI object detection systems may struggle to recognize individuals with darker skin tones (Wilson et al., 2019), LiDAR sensors can more easily detect larger objects (e.g., trucks) compared to smaller objects such as pedestrians (Zhang et al., 2020). Without the (timely) identification and inclusion of these traffic participants into the SDV’s calculations, the efficacy of our deliberately non-discriminatory decision-making model will be undermined. Lastly, while this model strives to incorporate numerous fundamental ethical principles simultaneously, it involves implicit tradeoffs since not all requirements/components are mutually compatible. To resolve these inconsistencies (e.g., equal risk distribution vs. minimization of aggregated risk), we stress the importance of including societal preferences or legal considerations as helpful ‘tie-breakers’ “to resolve fundamental ethical disagreements and thus garner public acceptability” (Evans, 2021; p.324).

? **Determining numerical figures and technical measures for value-laden terms.** As illustrated in Figure 25, a few terms³¹ that underlie each step of the proposed model are stated – on purpose – in an abstract manner at this stage. To turn this proposed model into practice, these terms will need to be concretized as numerical figures or derived from technical indicators in the future. These terms include precisely the fixation of *duties*, thresholds for collision probability (c_{max}) and estimated harm (h_{max}), the classification, hierarchy, and valence for certain traffic participant types (v_i), relevant *risk distribution principles* and their corresponding weighting factors (e.g., w_E and w_U). To approximate the numerical figures of some of these terms, the *technical measures/indicators in Table 17* may serve as a starting point. For example, to ensure compliance with duties such as “The lives of traffic participants must not be put in harm’s way” (Evans et al., 2020), these duties will need to be attached to specific risk figures (as illustrated exemplarily in Table 10), which, in turn, can be based on indicators such as time-to-collision or time-to-react (Wishart et al., 2020). In addition, the value-laden terms can draw on *established measures in other fields*. For example, utilized principles in other fields (e.g., in healthcare: treating people equally vs. maximizing total benefits) (Persad et al., 2009) can point to appropriate risk distribution principles and their relative importance (i.e., weighting). The determination of the two different threshold figures (i.e., c_{max} , h_{max}) could be based on established thresholds in other fields, such as individual dose limits of radiation exposure (Goodall, 2016). Furthermore, numerical figures of the value-laden terms could be approximated through *deliberations among experts* and informed by indicated preferences of other key stakeholders, such as the *broader society* (Poszler et al., 2024). For example, experts from Germany’s national ethics committee for automated and connected driving prohibited factoring in road users’ personal characteristics in SDV calculations (Lütge, 2017). This has implications for, amongst others, the ultimate determination of traffic participant types in that, for example, age shall not be a technical measure/indicator for this classification. The broader society can contribute to establishing numerical figures for the value-laden terms by indicating their preferences in empirical studies. For example, Meder et al. (2019) asked participants to state the minimum likelihood of colliding with a pedestrian (similar to c_{max}) in order to grant lane departure to an SDV in a dilemma situation. Similarly, to supplement our model, future research could ask participants to decide between various traffic scenarios that implicitly reflect different risk

³¹ The focus will be on the terms that carry ethical dimensions, i.e., are value-laden (such as maximum acceptable risk thresholds). By contrast, $c_{Ai,Ti}$ in step 1 will not be listed here since the calculation of collision probabilities is primarily an ‘objective’, technical matter as it depends, for example, on uncertainty estimations due to occluded areas (Nolte et al., 2018) that may emerge from existing obstacles in a traffic situation (ETSI, 2021).

distributions (e.g., Equality principle and Bayesian principle), thereby revealing the relative importance of particular distribution principles.

- ? **Empirical evaluations of the ethical decision-making process as a next step.** Assessing the actual impact, effectiveness, and admissibility of implementing the ethical principles and the decision-making process outlined here requires an empirical evaluation and testing of the entire model. Before real-world deployment, simulations could shed some light on the feasibility of implementing particular ethical theories, potential arising tradeoffs, or destructive outcomes (Hoffmann, 2021), for example, regarding fairness or safety levels (Eastman et al., 2023). These evaluations would, amongst others, test the degree to which the algorithm performs as intended or if key requirements are ultimately undermined and measure societal and individual repercussions (Awad et al., 2022), such as the level of traffic flow efficiency or the number of traffic-related casualties. To address one of the earlier mentioned potential technical issues, such simulations could also assess whether adequate time is available for computing ethical parameters in particular traffic situations (UNECE, 2021). Additionally, it could be investigated how particular SDVs' ethical decision-making processes function (or represent an improvement for traffic participants) in comparison to motion planning frameworks that car manufacturers actually use. Geisslinger et al. (2023a) can serve as an example of a simulation showing how risk distributions among traffic participants change when particular ethical theories are implemented into an SDV's trajectory planning algorithm. Similarly, scholars and practitioners could validate the values and decision-making model outlined here through corresponding simulations in the future.

4.5 Conclusion

Overall, this paper aims to establish an ethical decision-making process for SDVs in hazardous situations. In particular, expanding on existing approaches (e.g., Evans et al., 2020; Poszler et al., 2023; Robinson et al., 2021), this proposed model states where exactly which (ethical) theories and requirements may apply during the decision-making process and how they could be represented (as numerical figures) in an SDV's calculation. Although not exhaustive and resolute, this approach highlights some key considerations indicated by policymakers, standardization organizations, and scholars at this moment in time. Namely, the model utilizes overall risk (i.e., safety) as a central factor and takes into account context, reasonableness, responsibility, and the protection of vulnerable road users. Furthermore, the model allows the integration of a mix of ethical theories and societal values, and overall, provides a chronological order for particular decision-making steps while leaving room for future adjustments. The requirements and technical specifications provided in this article can serve as a register for

contemporary SDV developers, manufacturers, or so-called ‘value leads’ when eliciting relevant ethical considerations and turning them into concrete system features. Like any other proposed decision-making process (e.g., Evans et al., 2023), our model will require ongoing evaluation and adaptation to address emerging legal restrictions, ethical standards, or technical advancements.

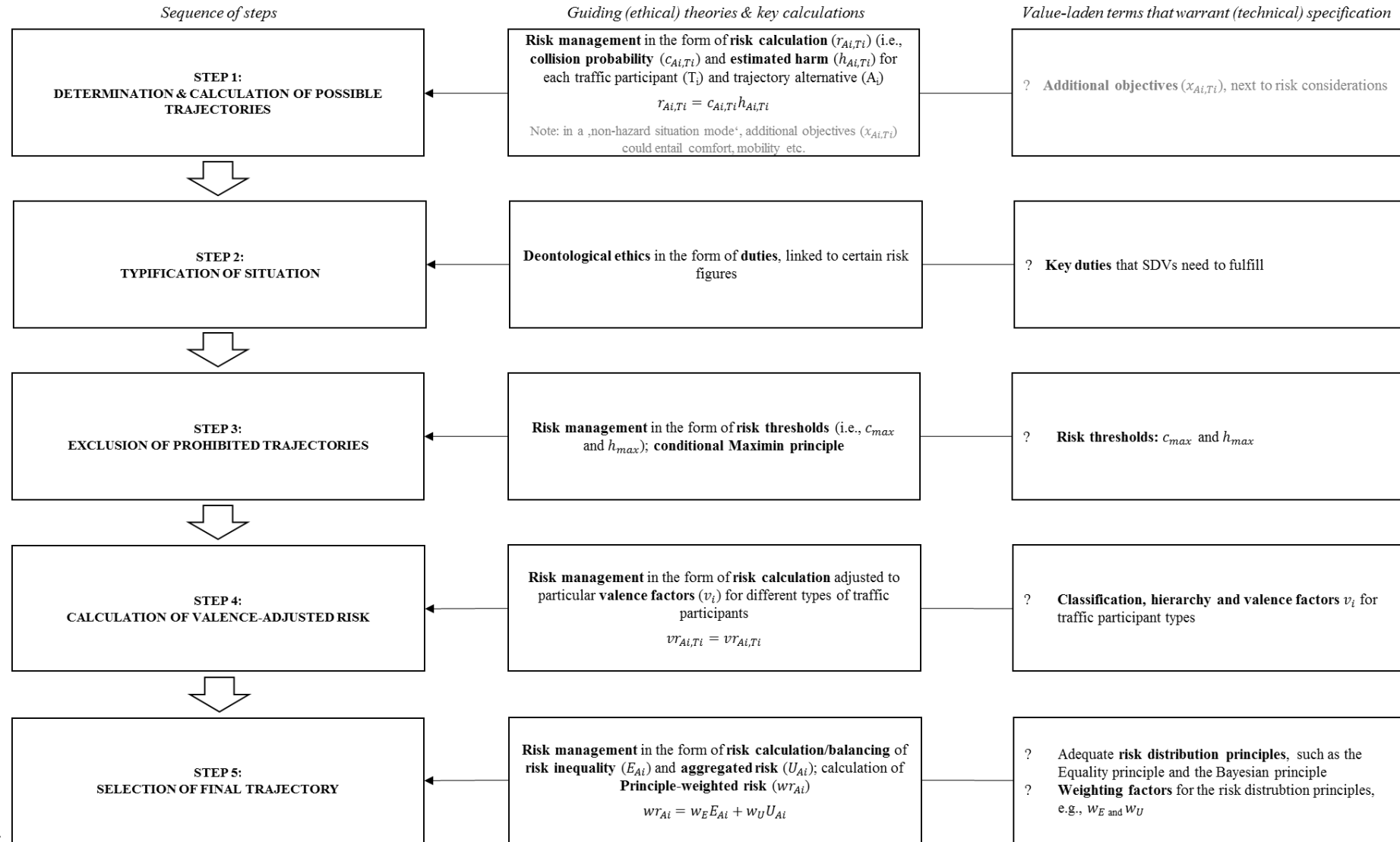


Figure 25: Proposed model summarizing the decision-making steps, guiding theories, underlying calculations, and terms that warrant (technical) specification

References

- Abdelhalim, A., & Abbas, M. (2022). A real-time safety-based optimal velocity model. *IEEE Open Journal of Intelligent Transportation Systems*, 3, 165-175. <https://doi.org/10.1109/OJITS.2022.3147744>
- Aksjonov, A., & Kyrki, V. (2023). A Safety-Critical Decision-Making and Control Framework Combining Machine-Learning-Based and Rule-Based Algorithms. *SAE International Journal of Vehicle Dynamics, Stability, and NVH*, 7 (3), 287-299. <https://doi.org/10.4271/10-07-03-0018>
- Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature*, 563(7729), 59-64. <https://doi.org/10.1038/s41586-018-0637-6>
- Awad, E., Levine, S., Anderson, M., Anderson, S. L., Conitzer, V., Crockett, M. J., Everett, J. A., Evgeniou, T., Gopnik, A., Jamison, J. C., Kim, T. W., Liao, S. M., Meyer, M. N., Mikhail, J., Opoku-Agyemang, K., Borg, J. S., Schroeder, J., Sinnott-Armstrong, W., Slavkovik, M., et al. (2022). Computational ethics. *Trends in Cognitive Sciences*, 26(5), 388-405. <https://doi.org/10.1016/j.tics.2022.02.009>
- Berkey, B. (2022). Autonomous Vehicles, Business Ethics, and Risk Distribution in Hybrid Traffic. In R. Jenkins, D. Cerny, & T. Hriebek (Eds.), *Autonomous Vehicle Ethics: The Trolley Problem and Beyond*, 210.
- Bin-Nun, A. Y., Derler, P., Mehdipour, N., & Tebbens, R. D. (2022). How should autonomous vehicles drive? Policy, methodological, and social considerations for designing a driver. *Humanities and social sciences communications*, 9(1), 1-13. <https://doi.org/10.1057/s41599-022-01286-2>
- Bundesministerium der Justiz (BMJ) (2023). *Straßenverkehrsgesetz (StVG)*. Retrieved from: <https://www.gesetze-im-internet.de/stvg/BJNR004370909.html>
- California Department of Motor Vehicles (DMV) (2022). *Article 3.7. Testing of Autonomous Vehicles*. Retrieved from: <https://www.dmv.ca.gov/portal/file/adopted-regulatory-text-pdf/>
- Coskun, S. (2021). Autonomous overtaking in highways: A receding horizon trajectory generator with embedded safety feature. *Engineering Science and Technology, an International Journal*, 24(5), 1049-1058.
- Dietrich, M., & Weisswange, T. H. (2019). Distributive justice as an ethical principle for autonomous vehicle behavior beyond hazard scenarios. *Ethics and Information Technology*, 21(3), 227-239. <https://doi.org/10.1007/s10676-019-09504-3>
- Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Springer Nature. <https://doi.org/10.1007/978-3-030-30371-6>
- Dyoub, A., Costantini, S., & Lisi, F. A. (2020). Logic programming and machine ethics. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2009.11186>
- D'Souza, J., Kim, J., & Pickering, J. (2023). Variable Trust Setting for Safe and Ethical Algorithms for Navigation of Autonomous Vehicles (C-NAV) on a Highway. In *Proceedings of the 20th International Conference on Informatics in Control, Automation and Robotics-Volume 1: ICINCO* (pp.88-96). *SciTePress*.

Eastman, B., Collins, S., Jones, R., Martin, J. J., Blumenthal, M. S., & Stanley, K. D. (2023). A Comparative Look at Various Countries' Legal Regimes Governing Automated Vehicles. *Journal of Law and Mobility*, 2023(1), 2.

ETSI (2018). *Intelligent Transport Systems (ITS); V2X Applications; Part 2: Intersection Collision Risk Warning (ICRW) application requirements specification*. Retrieved from: https://www.etsi.org/deliver/etsi_ts/101500_101599/10153902/01.01.01_60/ts_10153902v010101p.pdf

ETSI (2021a). *Intelligent Transport Systems (ITS); Vulnerable Road Users (VRU) awareness; Part 1: Use Cases definition; Release 2*. Retrieved from: https://www.etsi.org/deliver/etsi_tr/103300_103399/10330001/02.02.01_60/tr_10330001v020201p.pdf

ETSI (2021b). *Intelligent Transport Systems (ITS); Vulnerable Road Users (VRU) awareness; Part 2: Functional Architecture and Requirements definition; Release 2*. Retrieved from: https://www.etsi.org/deliver/etsi_ts/103300_103399/10330002/02.02.01_60/ts_10330002v020201p.pdf

European Commission (2020). *Ethics of connected and automated vehicles: Recommendations on road safety, privacy, fairness, explainability and responsibility*. Retrieved from: <https://op.europa.eu/en/publication-detail/-/publication/89624e2c-f98c-11ea-b44f-01aa75ed71a1/language-en/format-PDF/source-search>

European Commission (2021). *Regulatory framework proposal on artificial intelligence*. Retrieved from: <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>

European Commission (2022). *Commission implementing regulation (EU) 2022/1426*. Retrieved from: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R1426>

Evans, K. (2021). *The Implementation of Ethical Decision Procedures in Autonomous Systems: The Case of the Autonomous Vehicle*. Doctoral dissertation, Sorbonne université.

Evans, K., de Moura, N., Chauvier, S., Chatila, R., & Dogan, E. (2020). Ethical decision making in autonomous vehicles: The AV ethics project. *Science and Engineering Ethics*, 26(6), 3285-3312. <https://doi.org/10.1007/s11948-020-00272-8>

Evans, K., de Moura, N., Chauvier, S., & Chatila, R. (2023). Automated Driving Without Ethics: Meaning, Design and Real-World Implementation. In F. Fossa, & F. Cheli (Eds.), *Connected and Automated Vehicles: Integrating Engineering and Ethics* (pp.123-143). Springer Nature Switzerland.

Geisslinger, M., Poszler, F., Betz, J., Lütge, C., & Lienkamp, M. (2021). Autonomous driving ethics: From trolley problem to ethics of risk. *Philosophy & Technology*, 34(4), 1033-1055. <https://doi.org/10.1007/s13347-021-00449-4>

Geisslinger, M., Poszler, F., & Lienkamp, M. (2023a). An ethical trajectory planning algorithm for autonomous vehicles. *Nature Machine Intelligence*, 5(2), 137-144. <https://doi.org/10.1038/s42256-022-00607-z>

Geisslinger, M., Trauth, R., Kaljavesi, G., & Lienkamp, M. (2023b). Maximum acceptable risk as criterion for decision-making in autonomous vehicle trajectory planning. *IEEE Open Journal of Intelligent Transportation Systems*. <https://doi.org/10.1109/OJITS.2023.3298973>

Gogoll, J., & Müller, J. F. (2017). Autonomous cars: in favor of a mandatory ethics setting. *Science and engineering ethics*, 23, 681-700. <https://doi.org/10.1007/s11948-016-9806-x>

Goodall, N. J. (2016). Can you program ethics into a self-driving car?. *IEEE Spectrum*, 53(6), 28-58. <https://doi.org/10.1109/MSPEC.2016.7473149>

Government of Canada (2021). *Guidelines for testing automated driving systems in Canada*. Retrieved from: https://tc.canada.ca/en/road-transportation/innovative-technologies/connected-automated-vehicles/guidelines-testing-automated-driving-systems-canada#_Toc78892228

GOV.UK, Center for Data Ethics and Innovation (2022). *Responsible Innovation in Self-Driving Vehicles*. Retrieved from: <https://www.gov.uk/government/publications/responsible-innovation-in-self-driving-vehicles/responsible-innovation-in-self-driving-vehicles>

GOV.UK, Department for Transportation (2023). *The Highway Code*. Retrieved from: <https://www.gov.uk/guidance/the-highway-code/general-rules-techniques-and-advice-for-all-drivers-and-riders-103-to-158>

Hoffmann, C. H. N. (2021). On formal ethics versus inclusive moral deliberation. *AI and Ethics*, 1, 313-329. <https://doi.org/10.1007/s43681-021-00045-4>

Hübner, D., & White, L. (2018). Crash algorithms for autonomous cars: How the trolley problem can move us beyond harm minimisation. *Ethical Theory and Moral Practice*, 21(3), 685-698. <https://doi.org/10.1007/s10677-018-9910-x>

IEEE (2019). *ETHICALLY ALIGNED DESIGN – A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. Retrieved from: <https://standards.ieee.org/wp-content/uploads/import/documents/other/ead1e.pdf>

IEEE (2021). *IEEE Standard Model Process for Addressing Ethical Concerns during System Design*. Retrieved from: <https://engagestandards.ieee.org/ieee-7000-2021-for-systems-design-ethical-concerns.html>

IEEE (2022). *IEEE Standard for Assumptions in Safety-Related Models for Automated Driving Systems*. <https://doi.org/10.1109/IEEESTD.2022.9761121>

International Organization for Standardization (ISO) (2018). *ISO 26262-1:2018: Road vehicles — Functional safety — Part 1: Vocabulary*. Retrieved from: <https://www.iso.org/obp/ui/en/#iso:std:iso:26262:-1:ed-2:v1:en>

International Organization for Standardization (ISO) (2020a). *ISO 22078:2020: Intelligent transport systems — Bicyclist detection and collision mitigation systems (BDCMS) — Performance requirements and test procedures*. Retrieved from: <https://www.iso.org/obp/ui/en/#iso:std:iso:22078:ed-1:v1:en>

International Organization for Standardization (ISO) (2020b). *ISO/TR 4804:2020: Road vehicles — Safety and cybersecurity for automated driving systems — Design, verification and validation*. Retrieved from: <https://www.iso.org/obp/ui/en/#iso:std:iso:tr:4804:ed-1:v1:en>

International Organization for Standardization (ISO) (2021). *ISO/SAE PAS 22736:2021: Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles*. Retrieved from: <https://www.iso.org/obp/ui/fr/#iso:std:iso-sae:pas:22736:ed-1:v1:en>

International Organization for Standardization (ISO) (2022). *ISO 21448:2022: Road vehicles - Safety of the intended functionality*. Retrieved from: <https://www.iso.org/standard/77490.html>

International Organization for Standardization (ISO) (2023a). *ISO 23375:2023: Intelligent transport systems — Collision evasive lateral manoeuvre systems (CELM) — Requirements and test procedures*. Retrieved from: <https://www.iso.org/obp/ui/en/#iso:std:iso:23375:ed-1:v1:en>

International Organization for Standardization (ISO) (2023b). *ISO/DIS 39003:2023: Road Traffic Safety (RTS) – Guidance on ethical considerations relating to safety for autonomous vehicles*. Retrieved from: <https://www.iso.org/obp/ui/en/#iso:std:iso:39003:ed-1:v1:en>

Jacobs, N., & Hultgren, A. (2021). Why value sensitive design needs ethical commitments. *Ethics and information technology*, 23(1), 23-26. <https://doi.org/10.1007/s10676-018-9467-3>

Jensen, J. B. (2018). Self-driving but not self-regulating: The development of a legal framework to promote the safety of autonomous vehicles. *Washburn Law Journal*, 57(1), 579-611.

Justia US law (2022). *NV Rev Stat § 482A.044 (2022)*. Retrieved from: <https://law.justia.com/codes/nevada/2022/chapter-482a/statute-482a-044/>

Kauppinen, A. (2021). Who Should Bear the Risk When Self-Driving Vehicles Crash?. *Journal of Applied Philosophy*, 38(4), 630-645. <https://doi.org/10.1111/japp.12490>

Kriebitz, A., Max, R., & Lütge, C. (2022). The German Act on Autonomous Driving: why ethics still matters. *Philosophy & Technology*, 35(2), 1-13. <https://doi.org/10.1007/s13347-022-00526-2>

Lin, Y., & Althoff, M. (2023). CommonRoad-CriMe: A toolbox for criticality measures of autonomous vehicles. In *2023 IEEE Intelligent Vehicles Symposium (IV)* (pp.1-8). IEEE.

Liu, H. Y. (2017). Irresponsibilities, inequalities and injustice for autonomous vehicles. *Ethics and Information Technology*, 19(3), 193-207. <https://doi.org/10.1007/s10676-017-9436-2>

Liu, P., & Liu, J. (2021). Selfish or utilitarian automated vehicles? Deontological evaluation and public acceptance. *International Journal of Human-Computer Interaction*, 37(13), 1231-1242. <https://doi.org/10.1080/10447318.2021.1876357>

Lütge, C. (2017). The German ethics code for automated and connected driving. *Philosophy & Technology*, 30(4), 547-558. <https://doi.org/10.1007/s13347-017-0284-0>

Lütge, C., Poszler, F., Acosta, A. J., Danks, D., Gottehrer, G., Mihet-Popa, L., & Naseer, A. (2021). AI4People: Ethical Guidelines for the Automotive Sector–Fundamental Requirements and Practical Recommendations. *International Journal of Technoethics*, 12(1), 101-125. <https://doi.org/10.4018/IJT.20210101.0a2>

Meder, B., Fleischhut, N., Krumnau, N. C., & Waldmann, M. R. (2019). How should autonomous cars drive? A preference for defaults in moral judgments under risk and uncertainty. *Risk analysis*, 39(2), 295-314. <https://doi.org/10.1111/risa.13178>

Ministères Écologie Énergie Territoires (2022). *Safety validation of automated road transport systems: clarification through the analysis of accident data*. Retrieved from: https://www.ecologie.gouv.fr/sites/default/files/DGITM-Rapport_accidentelite-juillet_2022-EN.pdf

Mordue, G., Yeung, A., & Wu, F. (2020). The looming challenges of regulating high level autonomous vehicles. *Transportation research part A: policy and practice*, 132, 174-187. <https://doi.org/10.1016/j.tra.2019.11.007>

Munn, L. (2022). The uselessness of AI ethics. *AI and Ethics*, 3, 869-877. <https://doi.org/10.1007/s43681-022-00209-w>

Narayanan, A. (2019). Ethical judgement in intelligent control systems for autonomous vehicles. In *2019 Australian & New Zealand Control Conference (ANZCC)* (pp.231-236). IEEE. <https://doi.org/10.1109/ANZCC47194.2019.8945790>

National Association of City Transportation Officials (NACTO) (2016). *NACTO POLICY STATEMENT ON AUTOMATED VEHICLES*. Retrieved from: <https://nacto.org/wp-content/uploads/2016/06/NACTO-Policy-Automated-Vehicles-201606.pdf>

National Highway Traffic Safety Administration (NHTSA) (2022). *Occupant Protection for Vehicles With Automated Driving Systems*. Retrieved from: <https://www.nhtsa.gov/sites/nhtsa.gov/files/2022-03/Final-Rule-Occupant-Protection-Amendment-Automated-Vehicles.pdf>

National Transport Commission (NTC) (2022). *The regulatory framework for automated vehicles in Australia: Policy paper*. Retrieved from: <https://www.ntc.gov.au/sites/default/files/assets/files/NTC%20Policy%20Paper%20-%20regulatory%20framework%20for%20automated%20vehicles%20in%20Australia.pdf>

Németh, B. (2023). Coordinated Control Design for Ethical Maneuvering of Autonomous Vehicles. *Energies*, 16(10), 4254.

Nolte, M., Ernst, S., Richelmann, J., & Maurer, M. (2018). Representing the Unknown–Impact of Uncertainty on the Interaction between Decision Making and Trajectory Generation. In *2018 21st International Conference on Intelligent Transportation Systems* (pp.2412-2418). IEEE. <https://doi.org/10.1109/ITSC.2018.8569490>

Nyholm, S., & Smids, J. (2016). The ethics of accident-algorithms for self-driving cars: An applied trolley problem?. *Ethical theory and Moral Practice*, 19(5), 1275-1289. <https://doi.org/10.1007/s10677-016-9745-2>

Papadimitriou, E., Farah, H., van de Kaa, G., de Sio, F. S., Hagenzieker, M., & van Gelder, P. (2022). Towards common ethical and safe ‘behaviour’ standards for automated vehicles. *Accident Analysis & Prevention*, 174, 106724. <https://doi.org/10.1016/j.aap.2022.106724>

Persad, G., Wertheimer, A., & Emanuel, E. J. (2009). Principles for allocation of scarce medical interventions. *The Lancet*, 373(9661), 423-431. [https://doi.org/10.1016/S0140-6736\(09\)60137-9](https://doi.org/10.1016/S0140-6736(09)60137-9)

Poszler, F., Geisslinger, M., Betz, J., & Lütge, C. (2023). Applying ethical theories to the decision-making of self-driving vehicles: A systematic review and integration of the literature. *Technology in Society*, 75, 102350. <https://doi.org/10.1016/j.techsoc.2023.102350>

Poszler, F., Portmann, E., & Lütge, C. (2024). Formalizing ethical principles within AI systems: experts’ opinions on why (not) and how to do it. *AI and Ethics*, 1-29. <https://doi.org/10.1007/s43681-024-00425-6>

Rhim, J., Lee, G. B., & Lee, J. H. (2020). Human moral reasoning types in autonomous vehicle moral dilemma: a cross-cultural comparison of Korea and Canada. *Computers in Human Behavior*, 102, 39-56. <https://doi.org/10.1016/j.chb.2019.08.010>

Robinson, J., Smyth, J., Woodman, R., & Donzella, V. (2021). Ethical considerations and moral implications of autonomous vehicles and unavoidable collisions. *Theoretical Issues in Ergonomics Science*, 23(4), 435-452. <https://doi.org/10.1080/1463922X.2021.1978013>

SAE International (2021a). *Taxonomy and Definition of Safety Principles for Automated Driving System (ADS) – J3206_202107*. Retrieved from: https://www.sae.org/standards/content/j3206_202107/

SAE International (2021b). *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles – J3016_202104*. Retrieved from: https://www.sae.org/standards/content/j3206_202107/

Sütfeld, L. R., König, P., & Pipa, G. (2019). Towards a framework for ethical decision making in automated vehicles. *PsyArXiv preprint*. <https://doi.org/10.31234/osf.io/4duca>

Thornton, S. M., Pan, S., Erlien, S. M., & Gerdes, J. C. (2016). Incorporating ethical considerations into automated vehicle control. *IEEE Transactions on Intelligent Transportation Systems*, 18(6), 1429-1439. <https://doi.org/10.1109/TITS.2016.2609339>

Trauth, R., Moller, K., & Betz, J. (2023). Toward Safer Autonomous Vehicles: Occlusion-Aware Trajectory Planning to Minimize Risky Behavior. *IEEE Open Journal of Intelligent Transportation Systems*, 4, 929-942. <https://doi.org/10.1109/OJITS.2023.3336464>

United Nations Economic Commission for Europe (UNECE) (2021). *All you need to know about Automated Vehicles – Technical progress and regulatory activities*. Retrieved from: <https://unece.org/sites/default/files/2021-09/GRVA-11-26e.pdf>

United Nations Economic Commission for Europe (UNECE) (2022). *New Assessment/Test Method for Automated Driving (NATM) Guidelines for Validating Automated Driving System (ADS) – amendments to ECE/TRANS/WP.29/2022/58*. Retrieved from: https://unece.org/sites/default/files/2022-09/GRVA-14-16e_0.pdf

U.S. Department of Transportation (2021). *Automated Vehicles – Comprehensive Plan*. Retrieved from: https://www.transportation.gov/sites/dot.gov/files/2021-01/USDOT_AVCP.pdf

U.S. Department of Transportation & National Highway Traffic Safety Administration. (2016). *Federal Automated Vehicles Policy: Accelerating the Next Revolution in Roadway Safety*. Retrieved from: <https://www.transportation.gov/AV/federal-automated-vehicles-policy-september-2016>

Wang, H., Huang, Y., Khajepour, A., Cao, D., & Lv, C. (2020). Ethical decision-making platform in autonomous vehicles with lexicographic optimization based model predictive controller. *IEEE transactions on vehicular technology*, 69(8), 8164-8175. <https://doi.org/10.1109/TVT.2020.2996954>

Wanner, J., Herm, L. V., Langer, M., Imgrund, F., & Janiesch, C. (2020). A Moral Consensus Mechanism for Autonomous Driving: Towards a Law-compliant Basis of Logic Programming. In *Wirtschaftsinformatik (Zentrale Tracks)* (pp.17-32).

Westhofen, L., Neurohr, C., Koopmann, T., Butz, M., Schütt, B., Utesch, F., Neurohr, B., Gutenkunst, C., & Böde, E. (2023). Criticality metrics for automated driving: A review and suitability analysis of the state of the art. *Archives of Computational Methods in Engineering*, 30(1), 1-35. <https://doi.org/10.1007/s11831-022-09788-7>

Wilson, B., Hoffman, J., & Morgenstern, J. (2019). Predictive inequity in object detection. *arXiv preprint* arXiv:1902.11097. <https://doi.org/10.48550/arXiv.1902.11097>

Wishart, J., Como, S., Elli, M., Russo, B., Weast, J., Altekari, N., James, E., & Chen, Y. (2020). Driving safety performance assessment metrics for ads-equipped vehicles. *SAE International Journal of Advances and Current Practices in Mobility*, 2(5), 2881-2899. <https://doi.org/10.4271/2020-01-1206>

Woodgate, J., & Ajmeri, N. (2022). Principles for Macro Ethics of Sociotechnical Systems: Taxonomy and Future Directions. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2208.12616>

Zhang, F., Guan, C., Fang, J., Bai, S., Yang, R., Torr, P. H., & Prisacariu, V. (2020). Instance segmentation of lidar point clouds. In *2020 IEEE International Conference on Robotics and Automation (ICRA)* (pp.9448-9455). IEEE. <https://doi.org/10.1109/ICRA40945.2020.9196622>

Zhu, D. (2021). Research on Legal Risks and Legal Regulations of Autonomous Driving from the Perspective of Civil Law. *JL Pol'y & Globalization*, 108, 15-21. <https://doi.org/10.7176/JLPG/108-02>

Appendix of Essay III – List of relevant standards & regulatory documents

Number of standard	Title	Date of publication/update
ETSI TS 101 539	Intersection collision risk warning	June 2018
ISO 22078:2020	Intelligent transport systems — Bicyclist detection and collision mitigation systems (BDCMS) — Performance requirements and test procedures	February 2020
ETSI TS 103 300-2 V2.2.1	Intelligent Transport Systems (ITS); Vulnerable Road Users (VRU) awareness; Part 2: Functional Architecture and Requirements definition; Release 2	April 2021
ISO 21448:2022	Road vehicles – Safety of the intended functionality	June 2022
ISO 26262-1	Road vehicles — Functional safety — Part 1: Vocabulary	December 2018
SAE_J3016	Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles	April 2021
SAE J3206	Taxonomy and Definition of Safety Principles for Automated Driving System (ADS)	July 2021
IEEE P2846	P2846 – Assumptions for Models in Safety-Related Automated Vehicle Behavior	March 2022
ISO 23375:2023	Intelligent transport systems — Collision evasive lateral manoeuvre systems (CELM) — Performance requirements and test procedures	February 2023
ISO/TR 4804:2020	Road vehicles - Safety and cybersecurity for automated driving systems: Design, verification and validation	December 2020
ISO 39003	Road traffic safety (RTS) — Guidance on ethical considerations relating to safety for autonomous vehicles	July 2023
Country/issuer	Title of policy document/regulation	Date of publication/update
Germany – Federal Office of Justice	Road Traffic Act (Straßenverkehrsgesetz)	November 2023
USA – U.S. Department of Transportation & National Highway Traffic Safety Administration	Federal Automated Vehicles Policy: Accelerating the Next Revolution in Roadway Safety	September, 2016

Essay III: Ethical Decision-Making for Self-Driving Vehicles: A Proposed Model & List of Value-Laden Terms That Warrant (Technical) Specification

USA, California – California Department of Motor Vehicles	Article 3.7. Testing of Autonomous Vehicles	April 2022
USA, Nevada – Nevada Legislature	NV Rev Stat § 482A.044 (2022)	May 2022
USA – National Association of City Transportation Officials	Nacto Policy Statement on automated vehicles	June 2016
USA – National Highway Traffic Safety Administration	Occupant Protection for Vehicles With Automated Driving Systems	March 2022
European Commission – Directorate-General for Research and Innovation	Ethics of Connected and Automated Vehicles: Recommendations on road safety, privacy, fairness, explainability and responsibility	June 2020
European Commission	EUR-Lex - 32022R1426: Uniform procedures and technical specifications for the type-approval of the automated driving system (ADS) of fully automated vehicles	August 2022
United Nations Economic Commission for Europe	New Assessment/Test Method for Automated Driving (NATM) Guidelines for Validating Automated Driving System (ADS) – amendments to ECE/TRANS/WP.29/2022/58	September 2022
Canada – Transport Canada	Guidelines for testing automated driving systems in Canada	August 2021
United Kingdom – Center for Data Ethics and Innovation	Responsible Innovation in Self-Driving Vehicles	August 2022
United Kingdom – Department for Transportation	The Highway Code	September 2023
France – Ministères Écologie Énergie Territoires	Safety validation of automated road transport systems: clarification through the analysis of accident data	July 2022
Australia – National Transport Commission	The regulatory framework for automated vehicles in Australia: Policy paper	February 2022

Table 18: List of standards and regulatory documents in the field of autonomous driving ethics that were consulted in Essay III³²

³² Further lists of related standards can be found, for example, here: <https://www.connectedautomeddriving.eu/standards/standards-collection/>

5 | Discussion

As indicated in the Introduction and the Theoretical background, AI-enabled technologies take on ever more decisions that carry ethical dimensions, so they need to be programmed to make decisions in ‘ethically correct’ manners. Therefore, this dissertation centers on the topic of integrating ethical principles into AI systems. More specifically, the main goals of this dissertation were to investigate (1) how to practically integrate ethical principles into AI systems and (2) what societal implications would follow. As a reminder, the addressed research questions are:

- *Research question 1 (Practical investigation)*: How can ethical principles be integrated into AI systems?
- *Research question 2 (Societal investigation)*: What are societal implications of integrating ethical principles into AI systems?

To answer these questions, three individual studies were conducted. In this section, I briefly summarize these three separate essays. Afterward, overall implications for society (thereby answering Research question 2), practitioners (thereby answering Research question 1), and scholars will be drawn.

5.1 Summary of findings

The first essay addresses the research questions with a more general and broad perspective, meaning the practical and societal investigation is conducted by looking at the entire development process of AI systems and by not focusing on one specific technological application/AMA. By contrast, Essay II and Essay III are more specialized as they center on one concrete technological application (i.e., SDVs), on one precise activity in the design process (i.e., ‘ethics by design’), and on one part of the AI system (i.e., the trajectory planning algorithm). As illuminated in the following, the three essays build on each other in that identified open questions and findings of any preceding essays shape the inquiry of the consecutive essay(s), gradually adding pieces to the puzzle.

Essay I. This first study investigated the topic of encoding ethical principles into AI systems in an explorative manner and without focusing on a specific technological application but rather centered on artificial moral agents (AMAs) as a whole. Based on semi-structured interviews with twelve experts from philosophy, AI, and cognitive sciences, this study explored reasons for and against computational ethics and resulting AI systems. Findings suggest that indicated supporting

and opposing arguments can be clustered into practical, societal, and epistemic reasons, which need to be considered when engaging in computational ethics and developing AMAs. Despite the number of opposing arguments that call into question the *raison d'être* of computational ethics and AMAs, it is postulated that their (implicit) emergence is inevitable anyway, so efforts should be placed on how to develop them in a systematic and responsible manner. Correspondingly, this essay lists identified recommendations for companies' technological design and development, for industry's governance measures and future research directions. A model for facilitating computational ethics and AMAs was developed. This model depicts computational ethics as an integral part of the broader 'ethics in design' approach, which, in contrast, relates to the entire product/system lifecycle. It is further established that computational ethics require an 'ethical principles investigation' to assess what concrete ethical principles are integrated into the AI system. Reference points for the consulted ethical principles could be, for example, normative ethical theories, human values or legal standards, according to Essay I. Another indicated recommendation suggests that 'ethical principle investigations' and the subsequent development of AMAs should always be application-specific.

Essay II. Building on the insights of Essay I, this essay aimed to represent an exemplary 'ethical principles investigation' for one specific technological application (i.e., SDVs). Self-driving vehicles were selected as fitting technological applications since they will need to make ethical decisions that manufacturing companies have to program in advance. Based on a systematic review of the autonomous driving ethics literature indicated ethical theories and principles such as utilitarianism or Maximin were identified that could be applied to the logic of SDVs. These ethical theories were analyzed by summarizing corresponding social, moral/legal, and functional advantages and disadvantages mentioned in previous publications. The findings suggest that each ethical theory can be applied to the decision-making of SDVs, but neither of them is without negative ramifications. Instead, the adoption of 'mixed' algorithms is advocated, which combine elements from different ethical theories and considerations. A summarizing model was derived showing how each identified ethical theory could be technically implemented and integrated with each other to guide SDVs' decision-making. Overall, this essay aimed to set the groundwork for a reflected and effective integration of ethical principles into SDVs and deduce an updated research agenda. Namely, in the end, the essay stressed investigating, amongst others, the concrete chronological order with which an SDV would apply particular theories during its decision-making process. Thus, this essay is the preparatory work and an endorsement for Essay III.

Essay III. Based on insights from Essay II and requirements formulated in contemporary policy drafts, ethical guidelines, and (technical) standards, this essay proposed a five-step ethical

decision-making model for SDVs during hazardous situations in traffic. With the simultaneous consideration of ethical, legal, and engineering aspects, the essay hopes to establish a decision-making model for SDVs that warrants high levels of social acceptance, ethical acceptability, legal substantiation, and technical feasibility. As requested in the previous essay (i.e., Essay II), this model sets out a precise sequence of steps that an SDV could follow during motion planning. These steps span from determining and calculating possible trajectories to selecting the final trajectory, all of which are demonstrated using an imagined traffic scenario. For each step, the essay pointed out guiding (ethical) theories (e.g., the Maximin principle), key calculations, and a list of (value-laden) terms that need further investigation and technical specification in order to refine the model. Although not exhaustive nor resolute, this model provides – compared to the other two essays – the most concrete and illustrative answer to the question of how to integrate ethical principles into AI systems by setting out an explicit decision-making process for one exemplary technological application, namely, SDVs. Similarly, it seeks to represent an approach that puts a positive spin on the system’s resulting societal implications (i.e., moving away from ‘creation of harm & discrimination’ towards ‘consideration of well-being & fairness aspects’) (compare 5.2). In the future, assessing the actual impact, effectiveness, and admissibility of implementing the sketched decision-making model for SDVs requires empirical evaluations and testing (compare 5.4.2).

5.2 Implications for society

The individual essays of this dissertation illustrate that integrating ethical principles into AI systems hold significant implications for society (e.g., compare the societal and epistemic view introduced in Essay I, the social and moral/legal evaluation criteria indicated in Essay II, or the benefits stated in Essay III). This section summarizes key societal implications and, thereby, provides answers to the initially raised Research question 2. As illustrated in Figure 26, integrating ethical principles into AI systems can have positive societal implications (i.e., opportunities) and negative societal implications (i.e., risks).

In particular, integrating ethical principles into AI systems and thereby allowing technologies to take over ethical decisions from humans can entail the *avoidance of biases* and normative mistakes that humans, as ‘fallible beings’ and, due to their ‘bounded ethicality’, are prone to, which was introduced in Essay I. However, this process of de-biasing may only apply if AI systems are not a mirror of human biases. As illustrated in Essay II, the ability to manifest unbiased and rational decision-making within AI systems depends, amongst other factors, on the ethical theories consulted during the process of eliciting values and establishing ethical decision-making principles. For example, drawing on the theory of contractualism (i.e., by adopting the

veil of ignorance when bargaining what constitutes ‘correct’ decision-making of an AI) removes the possibility for humans to indicate self-interested preferences so that, rather mutually advantageous principles are enacted instead. On the other hand, as illustrated by Essay I and II, integrating ethical principles into AI systems can create the risk of reinforcing and *manifesting human biases* in these AI systems, for example, when encoding (discriminatory) preferences that were indicated in empirical studies. Similarly, Essay III acknowledges the issue of the ‘naturalistic fallacy’. This fallacy entails “inferring an ‘ought’ from an ‘is’” (Kim et al., 2019; p.1). In the context of this dissertation’s topic, it would entail determining which ethical principles to incorporate solely based on individuals’ values or preferences. Therefore, Essay I, II, and III suggest the simultaneous consideration of normative ethical theories, societal values, and legal requirements as one potential countermeasure.

Another positive effect is the *consideration of well-being and fairness aspects*. Namely, integrating ethical principles enables the acknowledgment of morally relevant factors in the decision process of AI systems, which, in turn, may allow the resulting outcomes to be ‘fair’ and beneficial for humans’ physical integrity. These societal benefits are, amongst others, also linked to the technology’s comparably high precision (e.g., in recognizing an obstacle in road traffic) and efficiency (e.g., in computing the best decision alternative/trajectory), as outlined in Essay I. Furthermore, the extent to which such positive outcomes are achieved depends on the adopted ethical theory. For example, as Essay II illustrates, a utilitarian programming of SDVs may create lower fatality rates in accident situations. As opposed to declarations made in the German Air Security Act of 2006, such general programming to reduce the number of personal injuries may be justifiable since individual victims are not known in advance (Lütge, 2024). Similarly, as Essay III suggests, ethical programming of SDVs would, for example, entail the special protection of vulnerable road users and, thereby, acknowledge that these traffic participants introduce less risk to traffic in the first place³³. By contrast, leaving up ethical decisions to AI systems can also generate (*physical*) *harm* to humans since AI systems are also error-prone, as revealed by Essay I. Similarly, Essay III argues that even with meticulous programming, SDVs may cause accidents due to, for example, flawed AI object detection systems or sensor malfunctions. As outlined in Essay II, integrating certain ethical theories into the functioning of SDVs can undermine the preservation of human rights or result in increased traffic casualties. For instance, incorporating the Maximin criterion as a risk distribution principle implies a preference for an infinite number of severe injuries over preventing a single death. Thus, this distribution principle may give undue

³³ This argument justifies the differential treatment of vulnerable road users compared to motorized road users, indicating that such algorithms are fair rather than biased (ISO, 2022).

weight to the moral claims/physical integrity of the worst-off while discriminating against (precautionary and) safer road users.

Furthermore, integrating ethical principles into AI systems may result in *humans' liberation from a few obligations* so that they have the spare capacity to pursue other tasks and duties. The underlying rationale is that doing so allows the AI systems to take over ethical decisions from humans. For example, as indicated in Essay I, utilizing AMAs in clinics could reduce the nurses' workload so they can dedicate more time to developing interpersonal connections with patients. Applying this idea to SDVs could mean that humans would no longer be considered drivers but rather passengers, allowing them to utilize their travel time for activities such as working, replying to emails, or reading. The magnitude of this opportunity relies on the SDV's capability to independently handle a wide range of traffic scenarios. This capability may vary depending on, amongst others, which ethical theory is integrated into the system, as stated in Essay II. For example, the evaluation in Essay II indicates adopting a utilitarian decision making yields a definite course of action for SDVs in any traffic situation. On the other hand, endowing AI systems with the capability to make ethical decisions may create *additional, new obligations for humans*. For example, as implied by Essay I, it may implicate AI's claim to moral patiency, so society needs to rethink and revise existing regulations by granting certain rights to AI systems.

In addition, having technologies support or directly suggest ethical decisions may lead to more *confidence of humans in their decisions*. As indicated in Essay I, such AI systems may serve as crosschecks or as consultants offering a second opinion that may coincide with or challenge the planned decisions of AI or AMA users, thereby validating their ethical decision-making. By contrast, integrating ethical principles into AI systems and introducing these systems as AMAs can result in humans' *overreliance on AI-suggested decisions* and in an uncritical/blind acceptance of provided solutions, which, in the worst case, might be unsound. A dramatic case of overreliance represents the fatal Uber crash in 2018, in which the SDV's human safety driver watched a TV show instead of monitoring imminent traffic events and the vehicle's operation (McGee, 2019).

Moreover, the conscious act of integrating ethical principles into AI systems and using AMAs for ethical decisions creates *transparency/traceability of decision processes*, producing more clarity in how a decision was reached. For example, Essay III outlines a precise decision-making model that clearly delineates the sequential steps, underlying ethical principles, and calculations to be adhered to for each decision, thereby directing the process through which SDVs determine their final decision in a (more) consistent/traceable manner. Furthermore, it can be argued that using AMAs for ethical decisions fosters transparency regarding the need to modify the decision-making procedure and the relevant adjustment skews. This is because underlying

algorithms, for one, work on symbolic reasoning and/or training data that can be inspected more easily than human deliberation processes. As indicated in Essay I, it is easier to identify and fix bugs in AI systems' logics than in humans. Additionally, adopting certain ethical theories contributes particularly well to creating transparency. For example, as Essay II suggests, implementing deontological ethics through well-defined rules such as "stay in lane" sets clear and traceable constraints to an SDV's decision-making. On the other hand, integrating ethical principles into AI systems and, as a result, allowing AI systems to make ethical decisions may generate *intransparency of decision processes and responsibility gaps*. The transparency issue especially relates to bottom-up or hybrid approaches to programming machines (Coeckelbergh, 2020). Essay I, therefore, suggests conducting a 'scope investigation' to assess and determine to what extent restrictions should be placed on the AMA behavior and use in this way. Furthermore, in contrast to the example of deontological ethics, adopting other ethical theories may lead to a lack of transparency. For instance, Essay II states adopting a utilitarian calculus, wherein the SDV aims for rigid optimization, could lead to unforeseeable vehicle behavior. Since neither the humans nor the AI system can fully bear responsibility for such shared and intransparent activities, Essay I emphasizes the risk of responsibility diffusion in this context. However, if manufacturers have put all their efforts into enhancing SDVs' safety to the highest degree achievable and the introduction of SDVs increases traffic safety overall, it could be argued that potential responsibility gaps are tolerable (Nyholm, 2023a).

By integrating ethical principles into AI systems, *humans can safeguard their control over AI behavior*. Essay I suggests that through computational ethics, individuals can proactively determine the values and morally relevant factors that underlie a technology's functionality, thereby restricting its capacity to independently develop its own decision-making logic. Essay II and III aim to serve as examples for human safeguards (e.g., system requirements and overarching decision-making models) designed to control – at least to some degree – the behavior of SDVs, thereby ensuring they operate in a way that benefits society. Therefore, similar to related approaches such as value-sensitive design, the sketched models in these essays help to proactively oppose technological determinism (Zuber et al., 2021), the idea "that technology develops as the sole result of an internal dynamic and then, unmediated by any other influence, molds society to fit its patterns" (Winner, 1986; p.21). On the other hand, this dissertation also stresses that when AI systems take on and execute ethical decisions, it can result in *technological determinism*. This situation might particularly arise when AI systems are developed using a bottom-up approach (i.e., machine learning), which empowers the system to identify decision rules that differ from the ethical principles/decision rules initially programmed by the system's designer, as outlined in

Essay I. Similarly, as Essay II suggests, using machine learning techniques, SDVs can independently generate moral rules that they then adhere to.

Another opportunity resulting from the integration of ethical principles into AI systems is the *social acceptance of corresponding AI* and the public's higher willingness to adopt pertinent technologies. Essay I indicates that this way, society may recognize the responsible design and development behind the technology, which, in turn, will lead them to trust the (benefits and benevolence of these) systems. To what extent AI systems are accepted by society may also depend on which ethical theories are embedded and to what extent they align with human values. For example, as depicted in Essay II, past studies have highlighted that individuals indicated preferences for utilitarian SDVs. Combining normative ethical theories with shared values (derived, for example, from public surveys) could facilitate the creation of social acceptance without committing to the naturalistic fallacy. Therefore, programming SDVs with regard to the recommendations in Essay II or III could potentially contribute to society's willingness to drive in these vehicles (Lütge, 2024). On the other hand, achieving social acceptance of such ethically informed systems is challenging due to *disagreements on the ethical baseline of AI*. Across Essay I, II, and III, it transpires that there is no ground truth or unanimous agreement regarding the precise ethical principles that should be operationalized within AI systems. Therefore, integrating ethical principles into AI systems could have negative effects on societal or individual acceptance levels for a particular AI system if they perceive that the 'wrong' principles have been adopted.

Lastly, addressing and engaging in the topic of integrating ethical principles into AI systems can have educational effects by providing *insights into human morality and ethics*. According to both Essay I and II, next to learning about the technical feasibility of computational ethics and subsequent consequences in the first place, its scientific inquiry generates knowledge about ethics and human morality overall. This is because this inquiry forces us to apply ethics/philosophy in a new and specific domain (i.e., machine ethics, autonomous driving ethics) and collect data about respective human judgment and preferences. For example, Essay III suggests undertaking a study to assess the relative importance of specific risk distribution principles for SDVs as a next step (see 4.4.2). Through such investigations, we can gain insights into individuals' preferences regarding distribution principles, potentially extending beyond the realm of risk allocations and autonomous driving. On the other side, integrating ethical principles into AI systems may *limit moral progress*, which is "the discovery and application of new values or sensitization to new sources of harm" (Frank, 2020; p.374). Namely, Essay I suggests focusing on the development and adoption of AMAs may induce technological solutionism, in which emphasis is placed on those principles that can be formalized while non-programmable ethical principles are neglected. This, in turn, may narrow moral progress, human morality, and ethics down to 'ethics-as-science'

(Schwarz, 2018). For instance, Essay II illustrates which specific ethical theories can be translated more effectively into algorithmic logic compared to others (e.g., see “(in)compatibility & (in)effectiveness of implementation” for utilitarianism and virtue ethics). The comparably easy implementation of a utilitarian calculus might skew AI systems toward adopting consequentialist decision-making processes in the future. In fact, the decision-making model for SDVs proposed in Essay III – similar to other existing models (Evans, 2021) – can be considered structurally consequentialist.

Overall, this dissertation’s identified opportunities and risks can be understood as opposite sides of the same coin (compare Figure 26). While many of the sketched opportunities and risks are not fundamentally new, AI raises these matters to the next level, in terms of their magnitude and scope (Lütge, 2024). Building upon Floridi et al.’s (2018) work, this dissertation contends that besides the inappropriate or excessive utilization of pertinent AI systems, the potential benefits or drawbacks of integrating ethical principles into AI systems and AMAs hinge on the particular ethical decision-making principles adopted and how they are practically implemented. This aligns with the idea that many negative societal implications can be avoided if AI systems were better designed (Spiekermann, 2023). Pertinent recommendations for practitioners derived from the three essays of this dissertation are summarized in the following section.

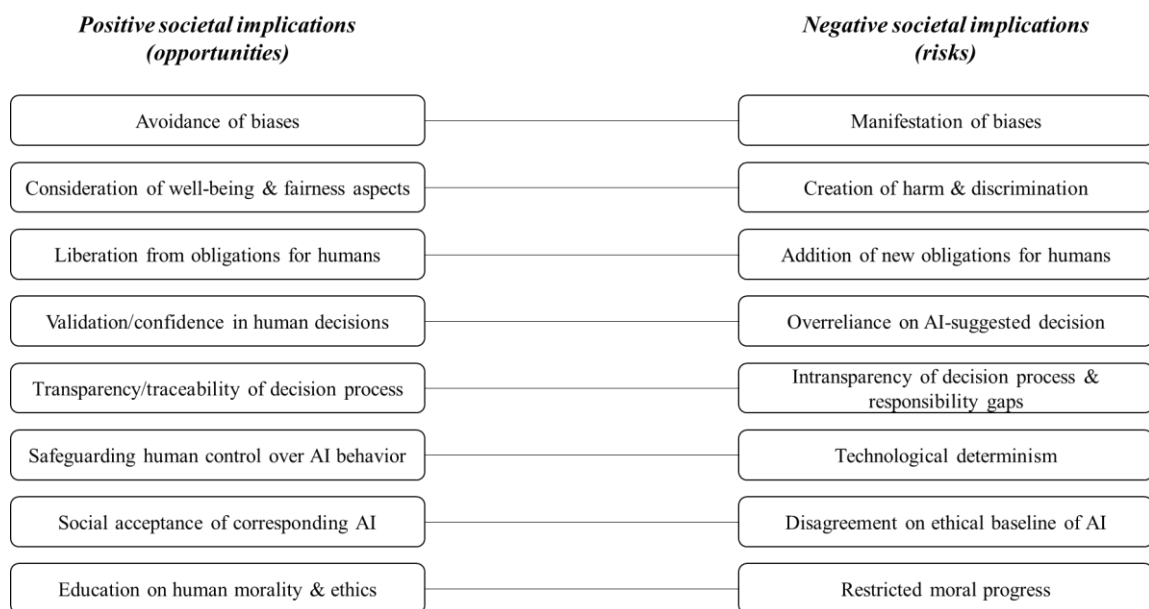


Figure 26: Overview of societal implications of integrating ethical principles into AI systems

5.3 Implications for practitioners

Each of the three essays in this dissertation highlights how practitioners (i.e., companies and policymakers) could practically facilitate, govern, and implement the integration of ethical principles into AI systems and the responsible development of AMAs (e.g., compare the recommendations for implementing computational ethics in Essay I, the functional evaluation criteria indicated in Essay II or the ethical decision-making model for SDVs proposed in Essay III). This section (see also Table 19) summarizes key practical outputs and takeaways and, thereby, provides answers to the initially raised Research question 1.

Essay I. The first essay provides an *overarching model illustrating how to facilitate computational ethics and artificial moral agents*. Similar to process models, this model visualizes an abstract workflow and key facilitating elements in a template form to develop ethically aligned AMAs (Prem, 2023). In particular, this model displays company-internal activities as well as external enablers and shows how computational ethics is an integral part of the overall ethical design of AMAs. In more detail, it is that part of the ethical design process of companies, which applies to the programming and underlying architecture of the algorithm itself (‘ethics by design’), while other activities rather relate to the ethical design and use of the resulting technology overall. As illustrated in Figure 9, ‘ethics by design’ requires, amongst others, an ‘ethical principles investigation’, which means that underlying reference points (e.g., certain ethical theories or AI principles) are examined for their applicability and their resulting impact when integrating them within an algorithm³⁴. Figure 9 also highlights that the development and governance of pertinent technology should be framed in ‘bigger picture’ reflections on potential ethical or societal implications.

Furthermore, the first essay lays out *recommendations for companies’ design and development* when engaging in the ethical design of AMAs and in the method of computational ethics. First, the overall ‘ethics in design’ approach should be organized in an iterative manner, including, for example, redress mechanisms to assess and adjust the technology’s proper/compliant behavior. Technically, companies can adopt three implementation strategies when integrating ethical principles within algorithms/AI systems, namely: via a top-down, bottom-up, or hybrid approach. Furthermore, key stakeholders involved in deciding underlying ethical principles are, for example, (ethical) experts and the broader society. Embedded ethical principles can be derived from, amongst others, a mix of normative theories, human values, legal

³⁴ Essay II serves as an example of what such an ‘ethical principles investigation’ could look like.

standards³⁵, company codes, or morally relevant contextual factors. Determining which particular ethical principles are eventually consulted should be application-specific, meaning ethical decision-making logics should not be copied and transferred between different technological applications. Other proposed restriction activities to computational ethics limit the autonomy and operation space of resulting AI systems. Lastly, transparent communication to users regarding the underlying (ethical) logic, functionalities, and limits of the AMAs is essential.

Essay I also derives *recommendations for policymakers*, helping them to govern computational ethics initiatives and the industry of pertinent technology. First, it is suggested to that ensure the market offers product diversity (with respect to the integrated ethical principles) and freedom of choice to consumers. This way, users could switch to alternative products if they disagree with the ethical decision logic of a particular AI system. Second, the entire industry developing and governing AMAs should adopt an interdisciplinary and multi-stakeholder approach, meaning cooperation partners stem from diverse disciplines such as philosophy, data science, and public policy. Similarly, policymakers are advised to initiate interdisciplinary educational programs by, for example, requiring engineering courses at universities that incorporate ethics. As a side effect, this fosters a shared language between engineers and ethicists, which is also seen as an important facilitator. After all, implementing ethics into algorithms/AI systems is futile if responsible parties (e.g., the programmers) are not trained or supported to do so. Third, policymakers should offer incentives and funding to companies and researchers pursuing computational ethics and the responsible development of AMAs.

As a last practical output, Essay I provides a *checklist that can be consulted when engaging in 'bigger picture' reflections and impact assessments* of AMAs. This will help practitioners to recognize and avoid imminent pitfalls (Prem, 2023). Namely, companies can consider the itemized practical, societal, and epistemic supporting or opposing arguments and examine to what extent they hold for their own technological applications and computational ethics approaches. Derived from Figure 6 and Figure 7, questions that companies would correspondingly ask themselves and check against include, for example:

- Are any particular biases (e.g., those of the AI system's designer or the project team) manifested?
- To what extent do our AI systems contribute positively to the users' confidence in their decisions?
- Where do our AI systems create accountability/responsibility gaps?

³⁵ Essay III serves as an example illustrating how normative theories, human values, established standards, and legal requirements can be consulted in the 'ethical principles investigation' when developing ethically aligned SDVs.

- Does the ethical design of our AI systems increase public acceptance and adoption levels?
- Are we neglecting ethical theories or AI principles/guidelines within the algorithms of our AI systems that are not as easily programmable?
- Do we have the relevant expertise and network to pursue the integration of ethical principles into our algorithms/AI systems?
- How can we control that our AI systems indeed realize embedded ethical principles?

Thus, the list of supporting and opposing arguments (see Figure 6 and Figure 7) can help companies to kick start ethical reflection processes, structure their risk assessments, ask the ‘right questions’, and reveal deficiencies in their design of AMAs that call for action. Similarly, policymakers could evaluate prevailing market trends and product advancements by drawing on the indicated supporting and opposing arguments as key considerations deserving careful scrutiny and oversight. The contemplation and evaluation of these considerations can ultimately underpin and legitimate certain (governance) activities by helping companies and policymakers identify areas that require deploying countermeasures and mitigation strategies. For instance, in line with IEEE Std 7000TM, it may serve helpful to discover potential personal value maxims or biases of the AI system’s designers or project team members that strongly shape the system’s operation (IEEE, 2022). To single out such biases, according to Essay I, companies could, for example, set up a council of ethicists who assess the ‘ethicality’ of an AMA. Furthermore, in cases where it is established that the necessary expertise is lacking to pursue the integration of ethical principles into algorithms/AI systems, companies can take corrective actions by instigating specific personnel or hiring strategies. At the same time, policymakers could initiate more funding opportunities for interdisciplinary cooperation projects.

Essay II. The second essay provides an *inventory of proposed ethical trajectory-planning algorithms and ethical theories that can be consulted in the programming of SDVs*. Namely, Table 3 summarizes all theories mentioned in past literature in this regard. Next to traditional ethical theories or considerations such as deontology, virtue ethics, consequentialism/utilitarianism, contractualism, metaethics, and descriptive ethics, these also include risk ethics. For each of these theories, it is indicated how they could be explicitly translated to the ethical decision-making/programming of SDVs. For example, deontological ethics could be manifested via rules (such as ‘an SDV must adhere to its designed path’) that direct the underlying logic of an SDV. In addition to these single theories, hybrid combinations, additional considerations/principles, and first elaborate decision approaches of ethical trajectory-planning algorithms are highlighted (see 3.3.4). Hybrid approaches constitute combinations of any previously mentioned individual theory (e.g., mix of traditional normative theories or a combination of traditional normative theories and risk ethics). Additional considerations/principles include non-arbitrary discrimination, the notion

of reasonableness, greatest equal chances/strict equality, randomization, prioritization, sensitivity to law and responsibility, situation-adjusted distribution, and distribution strategies from other fields such as medicine. First elaborate approaches that have been developed in this regard are the data theories method (Robinson et al., 2021), the ethical valence theory (Evans et al., 2020), or the expected moral value approach (Bhargava & Kim, 2017). In addition, Essay II itemizes existing proposed ethical trajectory-planning algorithms, such as the algorithm developed by Geisslinger et al. (2023) or Németh (2023). Derived from the comprehensive and structured literature review provided in Essay II, this list of stated ethical theories, considerations/principles, and elaborate approaches can serve practitioners as a reference guide or first indication of which ethical principles can be consulted when programming the decision-making of an SDV. By looking upon ethical theories beyond utilitarianism, virtue ethics, and duty ethics, this essay extends – compared to IEEE Std 7000TM (IEEE, 2021) – the catalog of theories that practitioners can refer to during AI system design. In addition, Essay II concludes that contemporary ethical trajectory-planning algorithms appear to follow the subsequent sequence of steps: (1) Assessing the present traffic scenario and all available trajectories, (2) Evaluating potential trajectories by weighing, restraining or prioritizing specific objectives and (3) Iteratively eliminating trajectories until the final ‘best’ action decision is selected. With this realization, the essay provides scholars or manufacturing companies with a basic structure of how SDVs could engage in decision-making processes³⁶.

Moreover, Essay II highlights the *advantages and disadvantages of applying particular ethical theories to the decision-making of SDVs* (see Table 4 and Table 5). Thus, in addition to the previously mentioned itemized list of ethical theories, etc., this dissertation offers a glimpse into their application’s related positive and negative implications. As indicated, these advantages and disadvantages can stem from a social, moral/legal, and functional perspective. The social perspective considers to what extent the particular ethical theory contributes to the transparency and predictability of the SDV behavior, the generation of optimal societal outcomes, and the compatibility with societal preferences/consensus. The moral/legal perspective looks at the extent to which the application of a particular ethical theory accords to the preservation of human rights, non-discrimination/the preservation of equality, and the extent to which it allows the consideration of relevant obligations and responsibility. The functional perspective analyzes to what extent the particular ethical theory is widely applicable, compatible, and effective in its implementation and corresponds with reality. This research output can be used in two different ways. First, the list of identified advantages and disadvantages may help companies and

³⁶ This basic structure, amongst others, serves as the foundation upon which the decision-making steps in Essay III originated.

policymakers decide which ethical theories to shortlist for integration into SDVs. By selecting any of the examined ethical theories in this essay, practitioners are given an initial insight into possible negative outcomes that need to be anticipated or that require the establishment of countermeasures, as well as positive outcomes that may even represent suitable solutions or countermeasures. For example, as stated in Table 5, implementing a utilitarian decision logic may result in ‘unfair’ risk distributions (e.g., higher safety levels for many traffic participants at the expense of one worst-off traffic participant). To counteract this outcome, the Maximin principle could be added to the SDV’s decision logic as it assures the protection of the most vulnerable traffic participant (compare Table 4). Second, practitioners can use the information and structure of Table 4 and Table 5 when engaging in their own ‘ethical principles investigation’ by, for example, extending the matrix with further ethical theories/principles and evaluating these with the indicated evaluation criteria.

Furthermore, Essay II carves out a *summarizing model for integrating ethical theories into the decision-making of SDVs* (see Figure 18). Risk ethics is an integral part of this model, as the nature of road traffic is defined by unpredictability rather than certainty. For example, it is uncertain whether a collision will occur or how traffic participants will act in the next few seconds. Accordingly, SDVs cannot rely on or compute outcomes as certainties but rather as risks (i.e., collision probability and estimated harm) that may result from selecting a particular trajectory. Furthermore, the model illustrates a combination of various ethical theories and postulates that applying a single theory is insufficient. Instead, it emphasizes the need for hybrid approaches to address tradeoffs, as demonstrated by one example in the previous paragraph. In addition, it is displayed abstractly how each ethical theory can be applied within an SDV’s decision-making (see Figure 18), for example:

- Derived from utilitarian conceptions, a cost-function can be built that weighs different theories, principles, and utilities (such as safety and comfort).
- Deontological ethics can be consulted as soft constraints to the SDV’s optimization process (e.g., in the form of rules or hierarchical orders).
- Contractualism in the form of the Maximin Principle can be integrated as an additional restriction to the SDV’s optimization function by instructing to bypass the most significant harm.
- Virtue ethics can be used to establish reinforcement learning signals or to determine the weight by which specific rules/decision principles are applied.
- Risk ethics can be adopted by calculating outcomes not as definite certainties but as uncertainties. It can also serve as a constraint in the form of thresholds for collision probabilities and estimated harm levels that are not to be exceeded.

- Insights from descriptive ethics can be implemented by utilizing indicated societal preferences to determine the weight or extent to which particular ethical theories, rules, or thresholds are considered.

Frameworks like these can “serve as a skeleton for addressing ethical aspects of an AI system” (Prem, 2023; p.705). With this particular model, practitioners (e.g., programmers and policymakers) are provided with a point of reference and a foundation on which they can start building and testing more concrete ethical decision-making logics of SDVs.

Essay III. Essay III is the most applied contribution and answer to Research question 1. Derived from the insights of the previous essay and propositions constructed in contemporary ethical guidelines, technical standards, and regulatory drafts, the third essay formulates a *list of crucial system requirements for SDVs’ ethical decision-making* (i.e., their trajectory planning) (see 4.2). Thus, the method of Essay III aligns with the IEEE 7000TM standard, which suggests always scrutinizing ethical principles “with a view to potential legal expectations and internationally applied ethical guidelines” (IEEE, 2021; p.15). Essay III itemizes relevant legal expectations and formulations set forth by regulatory entities (e.g., European Commission, 2022; GOV.UK 2022/2023), such as the necessity to classify road users and grant special protection to vulnerable road users. Correspondingly, Essay III (mathematically) illustrates how special protection should be provided for vulnerable road users such as pedestrians and cyclists. Other identified requirements include the centrality of risk considerations in the assessment and management of SDVs’ behavior (e.g., aiming to reach a ‘minimal risk condition’), thereby aligning with the findings of the second essay. In addition, Essay III postulates that SDVs need to adjust their calculations and decisions depending on the situation at hand, namely: While generally, SDVs are allowed/expected to consider a range of utilities such as comfort, mobility, safety, in ‘hazardous situations’, the only parameter to be considered is human life (i.e., physical integrity). Similar to Essay II, this essay suggests that the decision-making of SDVs should correspond to a mix of normative ethical theories and society’s values. Overall, since all proposed system requirements are linked to and endorsed by a registry of corresponding standards, regulations, and guidelines, Essay III can serve as a valuable resource for discovering fundamental documents and requirements that need to be consulted when engaging in value-sensitive design (of SDVs). Amongst others, these insights will help automotive companies anticipate and comply with essential obligations (that may be legally enacted in the future).

In addition, Essay III provides a *five-step ethical decision-making model for SDVs* (see 4.3 and Figure 25), which clarifies how the previously indicated system requirements can be turned into practice. Based on the three general steps identified in existing ethical trajectory-planning algorithms (according to Essay II), the model in Essay III depicts a detailed sequence of steps for

SDVs' decision-making in hazardous situations. This sequence of steps ranges from the determination and calculation of possible trajectories, the typification of the present traffic situation, the exclusion of prohibited trajectories, and the calculation of valence-adjusted risk to, ultimately, the selection of the final trajectory. For each step, guiding (ethical) theories are illustrated. For example, similar to Essay II, step 2 consults deontological ethics in the form of duties (e.g., 'The lives of traffic participants must not be put in harm's way'), or step 5 incorporates utilitarian considerations in the SDV's cost-function to minimize aggregated risk. In addition, each step is provided with underlying calculations, relevant terms, their definition, and corresponding technical measures. For instance, Table 17 in Essay III illustrates the term 'estimated harm' ($h_{Ai,Ti}$) can be defined as the "[s]everity of a collision in terms of the damage to the physical integrity of a human being (e.g., Evans, 2021; Geisslinger et al., 2021)". From a technical standpoint, estimated harm can be assessed by considering indicators such as Delta-v or the masses and impact angles of the colliding traffic participants, as outlined in Essay III. These definitions and technical measures/indicators stem from contemporary literature in the field of engineering autonomous driving (ethics). They can act as a reference to enhance the technical feasibility of computing ethical decision-making within SDVs. To add further practical relevance and applicability, the whole procedure is elaborated by running the five steps through an imagined, simplified traffic scenario. Overall, the decision-making model sketched in Essay III aims to establish theoretical foundations that not only take into account key aspects of existing law but also combine ethical and engineering guidelines/standards to develop an SDV's decision-making process that aims to be technically viable, legally permissible, ethically grounded and adaptable to societal values. This model could be seen as a provisional blueprint or skeleton that can be utilized, adjusted and tested by practitioners in the future.

Lastly, this essay summarizes a *list of value-laden terms that need to be concretized* in the future (see 4.4.2 and Figure 25). In particular, these are those terms that are indicated with exemplary numerical figures in the essay for explanatory purposes (e.g., " $w_E=0.5$ ") but are not to be taken at face value. In the future, practitioners and policymakers will need to dedicate time and effort toward the (technical) specification of these terms by investigating and determining the following questions:

- What additional utilities ($x_{Ai,Ti}$) – next to risk considerations – should SDVs take into account in 'non-hazard' situations?
- What essential duties must be breached for a situation to be characterized as 'hazardous'? What are the corresponding technical indicators?
- What are the exact thresholds for the maximum acceptable collision probability (c_{max}) and estimated harm (h_{max})?

- How many groups should traffic participants be classified into, and which specific groups? What are their corresponding valence factors (v_i)?
- What are (additional) relevant risk distribution principles (e.g., Equality principle and Bayesian principle)? How should these risk distribution principles be weighted (e.g., w_E and w_U)?

The further exploration and eventual execution of the outlined decision-making model demands addressing these questions and establishing numerical figures for these value-laden terms. Since they have yet to be sufficiently or precisely determined (e.g., in pertinent regulatory drafts or established standards), Essay III offers initial suggestions for suitable methods and benchmarks that could aid in approximating the concretization of these underlying terms. For example, to determine adequate risk distribution principles, responsible parties can draw on established principles in other fields such as medicine. Another method could involve employing empirical research to explore the weighting factors for the indicated risk distribution principles in accordance with societal preferences (see more in 5.4.2). Practitioners can use these potentially enlightening benchmarks as tools to concretize the compiled list of value-laden terms and, thereby, supplement the suggested decision-making model for SDVs in the future.

To sum up, the three essays in this dissertation show how ethical principles can be practically integrated into AI systems and which specific activities by companies and policymakers will contribute to this endeavor. Namely, this dissertation offers models, recommendations, checklists, inventories, evaluation criteria, and blueprints that responsible parties can refer to. Especially during periods when consensus on international regulations is still lacking, the ethical theories, AI principles/guidelines, or standards outlined here – potentially as precursors to legal requirements – could be considered by practitioners for orientation (Lütge, 2024)³⁷. In particular, this doctoral dissertation’s findings may help practitioners to engage in ethically aligned design activities and draft corresponding auditable repositories (i.e., ‘Case for Ethics’ as proposed in IEEE Std 7000™) (IEEE, 2021). Although Essay II and III focus on the specific case of SDVs, they nevertheless provide insights and practical outputs for integrating ethical principles into AI systems in general. Thus, the practical implications extend to policymakers and companies beyond the automotive sector.

³⁷ Even if pertinent regulations were to be established and enacted, this dissertation would not lose significance. After all, “there will always be aspects of AI systems that will not be covered by any law but need to be governed by ethics” (Vetter et al., 2023; p.35). In such cases, the sketched practical suggestions can be seen as complements, aiding in filling regulatory gaps through self-governance.

	Practical outputs	Key takeaways
Essay I	<ul style="list-style-type: none"> Overarching model illustrating how companies, policymakers, and scholars can facilitate computational ethics and artificial moral agents 	<ul style="list-style-type: none"> Central are ‘bigger picture’ ethical reflections/impact assessments ‘Ethics by design’ and computational ethics specifically are integral parts of the broader ‘ethics in design’ approach Computational ethics and the responsible development of AMAs require, amongst other investigations, an ‘ethical principles investigation’
	<ul style="list-style-type: none"> Recommendations for companies’ design and development (i.e., technical implementation methods, underlying reference points for ‘ethical principles investigations’, potential restrictions/requirements for the resulting technologies, organization of the workflow process) 	
	<ul style="list-style-type: none"> Recommendations for policymakers’/industry’s governance (i.e., provision of diversity in market supply, the establishment of interdisciplinary collaboration & education, and creation of incentives) 	
	<ul style="list-style-type: none"> Checklist for ‘bigger picture’ impact assessments (to spark necessary countermeasures) (i.e., indicated benefits and barriers from a practical, societal, and epistemic view) 	
Essay II	<ul style="list-style-type: none"> Reference guide: Inventory of proposed ethical trajectory-planning algorithms and ethical theories available for integration in SDV decision-making (e.g., deontology, consequentialism/utilitarianism, descriptive ethics, risk ethics, strict equality, randomization, data theories method) 	<ul style="list-style-type: none"> (Traditional) ethical theories can be integrated into the decision-making logics of SDVs No single ethical theory is sufficient on its own; mixed approaches are necessary Traditional ethical theories need to be combined with risk ethics Existing ethical trajectory-planning algorithms generally operate based on three steps in their decision-making process
	<ul style="list-style-type: none"> Criteria/matrix for the evaluation of applying certain ethical theories (i.e., indicated advantages and disadvantages from a social, moral/legal, and functional perspective) 	
	<ul style="list-style-type: none"> Summarizing model for integrating ethical theories into SDVs’ decision-making 	
Essay III	<ul style="list-style-type: none"> List of system requirements for SDVs facing hazardous traffic situations (supported by a register of corresponding standards, regulations, and ethical guidelines) 	<ul style="list-style-type: none"> SDVs need to adapt their decision-making to the situation’s criticality Special protection should be granted to vulnerable parties
	<ul style="list-style-type: none"> Blueprint: Explicit five-step ethical decision-making model for SDVs 	

Discussion

	(illustrating a precise sequence of steps, underlying calculations, key terms and their technical measures/indicators, as well as guiding (ethical) theories)	<ul style="list-style-type: none">Integrating ethical, legal, and engineering considerations is crucial in creating SDVs’ decision-making processes that strive to be technically viable, legally permissible, ethically grounded, and socially acceptableMany value-laden terms still require (technical) concretization
	<ul style="list-style-type: none">List of value-laden terms that warrant (technical) specification(e.g., duties, risk thresholds, valence factors for traffic participants, weighting factors for risk distribution principles)	

Table 19: Overview of the practical outputs and key takeaways of the three essays

5.4 Implications for scholars

In addition to the theoretical implications, limitations, and open questions listed in the individual essays, this section provides a concise overview of the dissertation's overall contribution to contemporary research (5.4.1), identifies its limitations, and suggests future research endeavors (5.4.2).

5.4.1 Contributions to research

While the approach employed in this dissertation generally draws on previous scholarly work and methods in the field of machine ethics and autonomous driving ethics, this section demonstrates the dissertation's contributions and expansions to this body of literature.

Offering practice-oriented machine ethics & autonomous driving ethics. As illustrated in the theoretical background of this dissertation (see 1.2), critics have highlighted deficiencies in current initiatives and academic outputs in AI ethics, noting their inability to offer meaningful, actionable, and effective approaches for developing ethical AI systems (Munn, 2022). To reiterate, although current AI principles and guidelines delineate essential values to contemplate when designing AI systems, they primarily do so in an abstract, hardly practical manner (Mittelstadt, 2019). For example, methods such as VSD do not provide a clear way of actually embedding values into a system's design. Also, standards such as "IEEE Std 7000™ do[es] not give specific guidance on the design of algorithms to apply ethical values such as fairness" (IEEE, 2021; p.12). Instead, these approaches focus on methods to elicit common values across stakeholder groups from which system requirements will be retrieved at some later point (Umbrello, 2019). Similarly, the autonomous driving ethics literature primarily consists of theoretical publications, with most articles examining and drawing analogies to the hypothetical and unrealistic Trolley Problem, as construed in Essay II (see Figure 21). By contrast, this dissertation demonstrates in a more applied manner how to turn AI ethics into practice (see 5.3). Namely, with its provided conceptual models (i.e., Figure 9, Figure 18, Figure 25) and the specified system requirements for SDVs outlined in Essay III, this dissertation delivers initial solutions on how to address and instantiate ethical principles within AI systems, mainly focusing on the values of 'safety' and 'fairness' within SDV decision-making. While the models put forward are still conceptual, they lay the theoretical groundwork for further investigations to tackle the technical challenge mentioned in Section 1.3.

Providing philosophical underpinnings & incorporating additional ethical and legal considerations. Contemporary methods, such as VSD run the risk of attending to mere stakeholder preferences rather than genuine moral values. To forgo the risk of the naturalistic fallacy, scholars have argued to complement the approach with an ethical theory (Jacobs &

Huldtgren, 2021). Therefore, IEEE 7000TM uses ethical theories as a lens to encounter the ramifications of technologies (i.e., benefits and risks). Namely, it is suggested that a utilitarian, virtue-ethical, and duty-ethical perspective be adopted to unveil relevant stakeholder values (Spiekermann, 2023). While VSD and IEEE 7000TM, in part, root their investigations in philosophical theories (Gerdes, 2022), they remain stakeholder-focused. By contrast, this dissertation adopts an approach informed by philosophy in the sense that traditional ethical theories are utilized as practical instruments or principles that AI systems (e.g., SDVs) draw on in their decision-making process. Ultimately, a “philosophical basis for value concepts” (IEEE, 2021; p.54), such as ‘safety’ or ‘fairness’ in the context of SDV decision-making, is established and operationalized. By doing so, this dissertation supports the claim that philosophy remains relevant and plays a vital role in today’s digitized world. Namely, ethics cannot only be helpful in guiding ex-post analyses but also for proactively instructing technological progress (Lütge, 2024). Therefore, this dissertation may inspire the scholarly community to further explore machine ethics with strong underpinnings in philosophy.

Moreover, the ethical theories here considered and analyzed extend beyond consequentialism, virtue ethics and deontological ethics. For example, Essay II (in Chapter 3 |) provides a more comprehensive analysis of ethical theories and principles that have been applied to the decision-making of SDVs and, thereby, sheds light on the advantages and disadvantages of adopting various theories in the context of autonomous driving. This scientific investigation seems timely and needed since “we are inundated with several ethical theories” (Segun, 2021; p.268) and, consequently, we are unsure about which one(s) to embrace (Dyoub et al., 2020). Furthermore, IEEE 7000TM states that the ethical design of AI systems should take into consideration existing ethical guidelines, standards, pertinent regulations, and so forth (Spiekermann, 2023). This dissertation does so by contemplating additional ethical guidelines, standards, and regulations that have been established in the field of autonomous driving (ethics) (see Essay III in Chapter 4 |). With the integration of diverse ethical theories, AI principles, standards, and pertinent regulations, this dissertation offers an initial answer to the normative challenge mentioned in Section 1.3.

Relevance for AI systems beyond SDVs & their decision-making. The majority of this dissertation (i.e., Chapter 3 | and Chapter 4 |) focuses on one specific technological application (i.e., SDVs) because “[d]etermining what is and is not ethically acceptable in a specific domain is a less daunting task than trying to devise a general theory of ethical and unethical behavior” (Anderson, 2011; p.25). Even though values such as ‘fairness’ can mean very different things from one context to another (Brey & Dainow, 2021; Spiekermann, 2023), scholars have argued that arising issues for which no profound body of literature or research exists yet can be informed

by previous discussions and case studies in related fields (e.g., McLennan et al., 2022). Therefore, the here-sketched results for the autonomous driving ethics literature could, in part, be informative and transferable to other sub-research domains of robot ethics and machine ethics (compare Sætra & Danaher, 2022), as indicated in Essay II (see 3.4.2). Additionally, the findings of this dissertation could extend beyond the decision-making of SDVs. For example, in addition to the software implications, the system requirements outlined in Essay III could be translated into implications for external hardware components of SDVs and, thus, inform scholars investigating SDV hardware features. For instance, the requirement to provide special protection for vulnerable traffic participants (see 4.2.3) could be adopted by installing signaling features that broadcast warnings to pedestrians in the event of an imminent crash. This way, proactive collision avoidance features (e.g., in the form of a collision avoidance algorithm) are complemented with passive avoidance features (e.g., in the form of warning signals) (ETSI, 2019). This example demonstrates that the findings of this dissertation may have implications beyond the academic community that focuses on the software or decision-making process of SDVs.

Summarizing, in line with the implications stated in the previous sections (5.2 and 5.3), this dissertation provides theoretical fundamentals on the practical implementation and societal implications of integrating ethical principles into AI systems. Thereby, it contributes to the scholarly work in machine ethics and computational ethics specifically, a topic that has received little attention in academic literature to this date (Segun, 2021). Next to the developed conceptual models of ‘how’ to integrate ethical principles into AI systems (e.g., compare Figure 9, Figure 18, and Figure 25), this dissertation adds a philosophically informed perspective to ongoing research debates (especially for the autonomous driving ethics literature).

5.4.2 Critical remarks, limitations, and future research agendas

Despite its contributions, this dissertation is not without limitations. This section lists a few precautions and caveats, raises open/unanswered questions, and points to fruitful research areas that can be explored in the future.

Missing delineation of specific SDV types. First, this dissertation focuses on ethical decision-making processes for SDVs at or above the automation Level 3, thereby not distinguishing between Levels 3, 4, and 5. In general, one could argue that as AI systems gain more autonomy, the activity of integrating ethical principles into them will become ever more critical (Wallach & Allen, 2009). After all, failing to establish boundaries of the freedom and creativity of AI systems and AMAs particularly could increase risks for society (Evans, 2021; Rossi & Mattei, 2019). Thus, with the increasing driving automation of SDVs beyond Level 3, the topic of this investigation will not lose its relevance. Instead, the generated results addressing Research

question 1 (i.e., the practical investigation) will increase in significance because highly or fully automated vehicles have even more freedom to decide (e.g., without human intervention and in more operational design domains). Some of the generated results addressing Research question 2 (i.e., the societal investigation) may be reinforced or undermined as the driving automation level of SDVs rises. For example, the nature and existence of responsibility gaps, one of the negative societal implications listed in Section 5.2, could shift with the introduction of Level-5 vehicles³⁸. To reiterate, while at Level 3, the human driver still has the option to take over driving tasks, this will not be the case for fully automated vehicles anymore (SAE International, 2021). Thus, for Level-5 vehicles, at least one participant – namely, the passenger inside the car – is excluded from the network of potentially liable parties. Although this network is thereby trimmed down at first sight, it does not necessarily mean it is less complex. Namely, it raises the question of who can fill this space to assume responsibility. This presents one issue of significant importance for SDVs beyond Level 3 and will need to be addressed in forthcoming research. Similarly, future investigations can analyze what other differences exist between Level 3, 4, and 5 cars with regard to the subject of this dissertation.

Second, this dissertation does not yet consider the vehicle-to-vehicle (V2V) connectivity of SDVs. Connected automated vehicles (CAVs) are expected to possess such communication capabilities (Hasibur Rahman & Abdel-Aty, 2021), potentially altering the implications and suggested implementation techniques. Therefore, it is again questionable to what extent the ethical decision-making model sketched here is equally applicable to CAVs, or whether it needs to be adjusted. For instance, assuming traffic participants are interconnected, the decision-making process described in Essay III might require an update to additionally collect and analyze the underlying logics of all other involved SDVs in real time, in order to derive an appropriate tactical decision. At this point, this is just one of many speculative ideas regarding how the generated findings differ for other types of SDVs, which future research could explore.

Limited stakeholder perspective: ethical acceptability vs. social acceptance. Compared to other approaches (e.g., VSD) (Aizenberg & Van Den Hoven, 2020), this dissertation does not conduct its investigation together with stakeholders, such as road users. Although studies examining users' preferences regarding the decision-making of SDVs in dilemma situations are referenced and analyzed (e.g., in Essay II), these preferences here constitute secondary data. However, "good governance of risky technology requires analyzing both social acceptance and

³⁸ It needs to be mentioned that although manufacturing companies such as Waymo or Tesla aim to produce cars of the highest level of automation (i.e., SAE Level 5) (Kosuru & Venkitaraman, 2023), it is questionable whether this will be successful and feasible any time soon, given prevailing technical obstacles and regulatory challenges (McKinsey & Company, 2024).

ethical acceptability” to be able to include both morally relevant aspects and stakeholder opinions (Taebi, 2017; p.1817). While this dissertation focuses on establishing ethically acceptable decisions of AMAs and SDVs specifically through conceptual philosophical contemplations (e.g., drawing on traditional moral theories), it falls short of addressing the social acceptance of sketched ideas with primary data. To complement and extend the findings of this dissertation, a next step could involve adopting ‘empirical ethics’ methods to “generate empirically informed ethical arguments” about how SDVs should decide in dilemmas or road traffic operations in general (Nyholm, 2023b; p.61) and what are corresponding societal preferences. After all, societal values are important underlying benchmarks for programming ethical AI, as mentioned throughout the essays of this dissertation. Some scholars even argue when no ground truth on ethical principles is available, we may draw on an “approximation as agreed upon by society” (Dwork et al., 2012; p.214). In any case, social agreement and generally acceptable compromises of the SDVs’ underlying functionality are essential for generating trust, acceptance, and adoption of SDVs (Bergmann, 2022). Hence, as stressed within the three individual essays, future research could build upon the findings of this dissertation with studies aiming to achieve social acceptance of particular ethical decision-making processes of AMAs and SDVs specifically. For example, user studies could help quantify the numerical figures of the value-laden terms outlined in Essay III (see 4.4.2). When setting up and conducting such experimental investigations or inclusive stakeholder debates, it will be important that they refrain from provoking “intuitive reactions to cartoonlike vignettes” (Nyholm, 2023b; p.64).

Empirical and technical validation of the proposed decision-making model. This dissertation touches upon all three investigation methodologies of VSD (i.e., conceptual, empirical, and technical) (Verbeek, 2011), with a primary focus on the conceptual part. However, thought experiments or imagined scenarios – such as the simplified traffic scenario in Essay III – “are not powerful enough for understanding large-scale interactions” among AI systems and their environment (Wallach & Allen, 2009; p.135). After all, even with a programmer’s utmost effort to develop an ‘ethical’ algorithm or decision-making logic (as outlined in this dissertation), biases could still creep in unintentionally, or in practice, the proposed model might fail to realize the values of ‘safety’ and ‘fairness’ as intended (van de Poel, 2020). Therefore, empirical and technical investigations are required to validate the here established models of integrating ethical principles into AI systems (e.g., Figure 9, Figure 18, and Figure 25). This entails explicitly, for example, testing the effectiveness of the proposed ethical decision-making principles and the decision-making model overall or assessing any interactions and tradeoffs that transpire in reality. For instance, empirical and technical studies could assess whether there are tradeoffs between certain ‘ethics in design’ and ‘ethics by design’ activities. For example, such studies could reveal

that privacy settings for SDVs, which prohibit the collection of specific data, hinder the vehicles' ability to make appropriate and reliable calculations because they lack access to the entirety of morally relevant factors. Other studies may discover that over time, traffic participants will adopt strategic behavior if they become aware that SDV decisions follow particular ethical principles (Yu et al., 2018). Overall, such assessments and validations need to be ongoing since it may only be possible to accurately understand an AI system's functionality and societal impact after it has been (widely) adopted (Peter et al., 2020; Wallach & Allen, 2009).

Normative nature of this dissertation? The aim of this dissertation was to investigate how ethical principles can be integrated into AI systems and explore subsequent societal implications. The societal implications are derived from qualitative expert interviews and a comprehensive review of the literature in the respective research domain. Thus, in general, a key objective of this dissertation was to provide a descriptive account of what has been previously said or written by others (e.g., compare Appendix B of Essay I – Frequency analyses of the experts' arguments & recommendations). However, because the findings concerning societal implications in this dissertation are grouped into positive and negative categories, the societal investigation (see Research question 2) conducted here is not purely descriptive. For example, experts' statements in Essay I are classified into reasons for and against AMAs and the method of computational ethics. In Essay II, the social and moral/legal arguments concerning applying particular ethical theories to the decision-making of SDVs are categorized into advantages and disadvantages. These categorizations entail a valuation of the obtained results and might subtly reinforce the normative obligation to partake in actions that lead to the here-sketched beneficial outcomes, while avoiding those associated with adverse effects on society.

Similarly, the results derived as a response to Research question 1 (i.e., the practical investigation) are framed as recommendations, blueprints, etc., that could be adopted by practitioners (see 5.3). Since this dissertation combines ideas from traditional ethical theories, other morally relevant considerations, and guidelines to establish these recommendations or blueprints, it may be argued that the formulated ethical decision-making logic for SDVs approximates ethical acceptability and some form of "a general, higher-order ideal" (Nyholm, 2023b; p.68). Nevertheless, it must be stressed that there is no objectivity in ethics (LaCroix & Luccioni, 2022). Therefore, the word 'can' in the guiding research question is vital and emphasizes that the sketched practical implications are not to be treated as normative prescriptions, as also indicated in the discussion sections of the three individual essays (e.g., see 3.4.2). Even if forthcoming empirical and technical research were to discover that, for instance, the decision-making model for SDVs proposed in Essay III can be implemented in practice, it does not necessarily mean that the model should be adopted. After all, 'can' does not imply

‘ought’: “[I]f ‘X can do Y’ is true then ‘X ought to do Y’ may be true or may be false” (Collingridge, 1977; p.350; Hampshire et al., 1951). Thus, similar to prior efforts (e.g., Evans 2021), this dissertation should not be seen as an ultimate normative solution to decision-making in SDVs but rather as laying a first foundation for the design of ethical SDVs (and AMAs). If (ever³⁹) an ultimate answer to artificial morality for any specific AMA is established, another question to be tackled arises as to whether it should be standardized and mandatory for all AMAs of the same sort.

³⁹ It could be argued that morality evolves gradually, “with refinements and improvements over time, and without the expectation that there is a final state of perfect knowledge” (IEEE, 2021; p.61; Legg & Hookway, 2021).

References

- Aizenberg, E., & Van Den Hoven, J. (2020). Designing for human rights in AI. *Big Data & Society*, 7(2), 2053951720949566. <https://doi.org/10.1177/2053951720949566>
- Anderson, S. L. (2011). Machine metaethics. In M. Anderson, & S. L. Anderson (Eds.), *Machine ethics* (pp. 21-27). Cambridge University Press.
- Bergmann, L. T. (2022). Ethical Issues in Automated Driving – Opportunities, Dangers, and Obligations. In A. Riener, M. Jeon, & I. Alvarez (Eds.), *User Experience Design in the Era of Automated Driving* (pp.99-121). Springer. https://doi.org/10.1007/978-3-030-77726-5_5
- Bhargava, V., & Kim, T. W. (2017). Autonomous vehicles and moral uncertainty. In P. Lin, R. Jenkins, & K. Abney (Eds.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence* (pp.5-19). Oxford University Press. <https://doi.org/10.1093/oso/9780190652951.003.0001>
- Brey, P., & Dainow, B. (2021). Ethics by design and ethics of use in AI and robotics. The SIENNA project-Stakeholder-informed ethics for new technologies with high socioeconomic and human rights impact. *SIENNA*. Retrieved from: https://sienna-project.eu/digitalAssets/915/c_915554-l_1-k_sienna-ethics-by-design-and-ethics-of-use.pdf
- Coeckelbergh, M. (2020). *AI ethics*. The MIT Press.
- Collingridge, D. G. (1977). 'Ought-Implies-Can' and Hume's Rule. *Philosophy*, 52(201), 348-351. <https://doi.org/10.1017/S0031819100027194>
- Dwork, C., Moritz, H., Toniann, P., Reingold, T., & Rich, Z. S. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp.214–226). Association for Computing Machinery. <https://doi.org/10.1145/2090236.2090255>
- Dyoub, A., Costantini, S., Lisi, F. A., & Letteri, I. (2020). Logic-based Machine Learning for Transparent Ethical Agents. In *CILC* (pp. 169-183).
- ETSI (2019). *Intelligent Transport System (ITS); Vulnerable Road Users (VRU) awareness; Part 1: Use Cases definition; Release 2*. Retrieved from: https://www.etsi.org/deliver/etsi_tr/103300_103399/10330001/02.01.01_60/tr_10330001v020101p.pdf
- European Commission (2022). *Commission implementing regulation (EU) 2022/1426*. Retrieved from: <https://eurlex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R1426>
- Evans, K. (2021). *The Implementation of Ethical Decision Procedures in Autonomous Systems: The Case of the Autonomous Vehicle*. Doctoral dissertation, Sorbonne université.
- Evans, K., de Moura, N., Chauvier, S., Chatila, R., & Dogan, E. (2020). Ethical decision making in autonomous vehicles: The AV ethics project. *Science and Engineering Ethics*, 26(6), 3285-3312. <https://doi.org/10.1007/s11948-020-00272-8>
- Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Lütge, C., Madelin, R., Pagallo, U., Rossi, F., et al. (2018). An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28, 689-707. <https://doi.org/10.1007/s11023-018-9482-5>

Frank, L. E. (2020). What do we have to lose? Offloading through moral technologies: moral struggle and progress. *Science and Engineering Ethics*, 26(1), 369-385. <https://doi.org/10.1007/s11948-019-00099-y>

Geisslinger, M., Poszler, F., Betz, J., Lütge, C., & Lienkamp, M. (2021). Autonomous driving ethics: From trolley problem to ethics of risk. *Philosophy & Technology*, 34(4), 1033-1055. <https://doi.org/10.1007/s13347-021-00449-4>

Geisslinger, M., Poszler, F., & Lienkamp, M. (2023). An ethical trajectory planning algorithm for autonomous vehicles. *Nature Machine Intelligence*, 5(2), 137-144. <https://doi.org/10.1038/s42256-022-00607-z>

Gerdes, A. (2022). A participatory data-centric approach to AI Ethics by Design. *Applied Artificial Intelligence*, 36(1), 2009222. <https://doi.org/10.1080/08839514.2021.2009222>

GOV.UK, Center for Data Ethics and Innovation (2022). *Responsible Innovation in Self-Driving Vehicles*. Retrieved from: <https://www.gov.uk/government/publications/responsible-innovation-in-self-driving-vehicles/responsible-innovation-in-self-driving-vehicles>

GOV.UK, Department for Transportation (2023). *The Highway Code*. Retrieved from: <https://www.gov.uk/guidance/the-highway-code/general-rules-techniques-and-advice-for-all-drivers-and-riders-103-to-158>

Hampshire, S., MacLagan, W. G., & Hare, R. M. (1951). Symposium: Freedom of the will. *Proceedings of the Aristotelian Society*, Supplementary Volumes, 25, 161-216.

Hasibur Rahman, M., & Abdel-Aty, M. (2021). Application of connected and automated vehicles in a large-scale network by considering vehicle-to-vehicle and vehicle-to-infrastructure technology. *Transportation research record*, 2675(1), 93-113. <https://doi.org/10.1177/036119812096310>

IEEE (2021). *IEEE Std 7000™-2021: IEEE Standard Model Process for Addressing Ethical Concerns during System Design*. Retrieved from: <https://engagestandards.ieee.org/ieee-7000-2021-for-systems-design-ethical-concerns.html>

ISO (2021). *ISO/SAE PAS 22736:2021 - Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles*. Retrieved from: <https://www.iso.org/standard/73766.html>

ISO (2022). *ISO/IEC 22989:2022 - Information technology, Artificial intelligence. Artificial intelligence concepts and terminology*. Retrieved from: <https://www.iso.org/standard/74296.html>

Jacobs, N., & Hultgren, A. (2021). Why value sensitive design needs ethical commitments. *Ethics and information technology*, 23(1), 23-26. <https://doi.org/10.1007/s10676-018-9467-3>

Kim, T. W., Donaldson, T., & Hooker, J. (2019). Grounding value alignment with ethical principles. *arXiv preprint arXiv:1907.05447*. <https://doi.org/10.48550/arXiv.1907.05447>

Kosuru, V. S. R., & Venkitaraman, A. K. (2023). Advancements and challenges in achieving fully autonomous self-driving vehicles. *World Journal of Advanced Research and Reviews*, 18(1), 161-167. <https://doi.org/10.30574/wjarr.2023.18.1.0568>

LaCroix, T., & Luccioni, A. S. (2022). Metaethical perspectives on 'Benchmarking' AI ethics. *arXiv preprint arXiv:2204.05151*. <https://doi.org/10.48550/arXiv.2204.05151>

Legg, C. & Hookway, C. (2021). Pragmatism. *The Stanford Encyclopedia of Philosophy*. Retrieved from: <https://plato.stanford.edu/entries/pragmatism/>

Lütge, C. (2024). *Wirtschaftsethik in realistischer Perspektive*. Mohr Siebeck. <https://doi.org/10.1628/978-3-16-162808-5>

McGee, P. (2019). Uber back-up driver faulted in fatal autonomous car crash. *Financial Times*. Retrieved from: <https://www.ft.com/content/6d0c5544-0afb-11ea-bb52-34c8d9dc6d84>

McKinsey & Company (2024). *Autonomous vehicles moving forward: Perspectives from industry leaders*. Retrieved from: <https://www.mckinsey.com/features/mckinsey-center-for-future-mobility/our-insights/autonomous-vehicles-moving-forward-perspectives-from-industry-leaders>

McLennan, S., Fiske, A., Tigard, D., Müller, R., Haddadin, S., & Buyx, A. (2022). Embedded ethics: a proposal for integrating ethics into the development of medical AI. *BMC Medical Ethics*, 23(1), 6. <https://doi.org/10.1186/s12910-022-00746-3>

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature machine intelligence*, 1(11), 501-507. <https://doi.org/10.1038/s42256-019-0114-4>

Munn, L. (2022). The uselessness of AI ethics. *AI and Ethics*, 3, 869-877. <https://doi.org/10.1007/s43681-022-00209-w>

Németh, B. (2023). Coordinated Control Design for Ethical Maneuvering of Autonomous Vehicles. *Energies*, 16(10), 4254. <https://doi.org/10.3390/en16104254>

Nyholm, S. (2023a). Minding the Gap (s): Different Kinds of Responsibility Gaps Related to Autonomous Vehicles and How to Fill Them. In In F. Fossa, & F. Cheli (Eds.), *Connected and Automated Vehicles: Integrating Engineering and Ethics* (pp. 1-18). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-39991-6_1

Nyholm, S. (2023b). *This is technology ethics: An introduction*. John Wiley & Sons.

Peters, D., Vold, K., Robinson, D., & Calvo, R. A. (2020). Responsible AI—two frameworks for ethical design practice. *IEEE Transactions on Technology and Society*, 1(1), 34-47. <https://doi.org/10.1109/TTS.2020.2974991>

Prem, E. (2023). From ethical AI frameworks to tools: a review of approaches. *AI and Ethics*, 3(3), 699-716. <https://doi.org/10.1007/s43681-023-00258-9>

Robinson, J., Smyth, J., Woodman, R., & Donzella, V. (2021). Ethical considerations and moral implications of autonomous vehicles and unavoidable collisions. *Theoretical Issues in Ergonomics Science*, 23(4), 435-452. <https://doi.org/10.1080/1463922X.2021.1978013>

Rossi, F., & Mattei, N. (2019). Building ethically bounded AI. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33) (pp. 9785-9789). <https://doi.org/10.1609/aaai.v33i01.33019785>

SAE International (2021). *Surface vehicle recommended practice: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles (J3016)*. Retrieved from: https://www.sae.org/standards/content/j3016_202104/

Sætra, H. S., & Danaher, J. (2022). To each technology its own ethics: The problem of ethical proliferation. *Philosophy & Technology*, 35(4), 93. <https://doi.org/10.1007/s13347-022-00591-7>

Schwarz, E. (2018). Technology and moral vacuums in just war theorising. *Journal of International Political Theory*, 14(3), 280-298. <https://doi.org/10.1177/1755088217750689>

Segun, S. T. (2021). From machine ethics to computational ethics. *AI & SOCIETY*, 36(1), 263-276. <https://doi.org/10.1007/s00146-020-01010-1>

Spiekermann, S. (2023). *Value-Based Engineering: A Guide to Building Ethical Technology for Humanity*. De Gruyter. <https://doi.org/10.1515/9783110793383>

Taebi, B. (2017). Bridging the gap between social acceptance and ethical acceptability. *Risk analysis*, 37(10), 1817-1827. <https://doi.org/10.1111/risa.12734>

Umbrello, S. (2019). Beneficial artificial intelligence coordination by means of a value sensitive design approach. *Big Data and Cognitive Computing*, 3(1), 5. <https://doi.org/10.3390/bdcc3010005>

Van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. *Minds and Machines*, 30(3), 385-409. <https://doi.org/10.1007/s11023-020-09537-4>

Verbeek, P. P. (2011). *Moralizing technology: Understanding and designing the morality of things*. University of Chicago press.

Vetter, D., Amann, J., Bruneault, F., Coffee, M., Düdler, B., Gallucci, A., ... & Z-Inspection® initiative (2023). Lessons learned from assessing trustworthy AI in practice. *Digital Society*, 2(3), 35. <https://doi.org/10.1007/s44206-023-00063-1>

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Winner, L. (1986). *The whale and the reactor: A search for limits in an age of high technology*. University of Chicago Press.

Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., & Yang, Q. (2018). Building ethics into artificial intelligence. *arXiv preprint arXiv:1812.02953*. <https://doi.org/10.48550/arXiv.1812.02953>

Zuber, N., Gogoll, J., Kacianka, S., Pretschner, A., & Nida-Rümelin, J. (2022). Empowered and embedded: ethics and agile processes. *Humanities and Social Sciences Communications*, 9(1), 1-13. <https://doi.org/10.1057/s41599-022-01206-4>

6 | Conclusion


Many AI systems today, and even more so in the future, will inevitably take on the role of ‘implicit ethical agents’ by operating in environments where the decisions they make carry ethical dimensions. It is in our hands to turn these systems into ‘explicit ethical agents’ by equipping them with artificial morality, that is, a decisional procedure which allows AI systems to make explicit use of ethical considerations in their decision-making. By doing so, we can proactively shape the development of AI systems and generate artificial moral agents that promote positive values, such as fairness, and hinder the manifestation of negative values, such as discrimination. Nevertheless, the translation of high-level values into concrete decision-making principles or design requirements for AI systems remains an ongoing challenge, with effective, admissible solutions yet to be found. With its three individual essays, this doctoral dissertation addresses this challenge by offering practice-oriented groundwork on how to integrate ethical principles into AI systems, specifically self-driving vehicles, and sensitizes the need for this endeavor by illustrating societal implications. Therefore, this dissertation contributes to the academic fields of AI ethics, machine ethics, and autonomous driving ethics. It also showcases to practitioners how to develop ‘ethical’ artificial moral agents, using the exemplary case of self-driving vehicles. Whether out of self-interest or because they are legally obligated to, individuals may utilize and expand upon the findings of this dissertation to develop and enforce responsible ‘ethics by design’ practices.

Appendix

Appendix A: Reference & copyright information by the publisher for the first essay (Essay I, Chapter 2)


Poszler, F., Portmann, E., & Lütge, C. (2024). Formalizing ethical principles within AI systems: experts' opinions on why (not) and how to do it. *AI and Ethics*, 1-29.
<https://doi.org/10.1007/s43681-024-00425-6>

Acknowledgement: Reproduced with permission from Springer Nature



Mo 19.02.2024 14:21
[Journalpermissions <journalpermissions@springernature.com>](mailto:journalpermissions@springernature.com)
RE: Permission letter for article inclusion in my doctoral disseration

An ☐ Franziska Poszler

 Zur Nachverfolgung. Erledigt am Montag, 19. Februar 2024.
Klicken Sie hier, um Bilder herunterzuladen. Um den Datenschutz zu erhöhen, hat Outlook den automatischen Download von Bildern in dieser Nachricht verhindert.

Dear Franziska,

Thank you for your recent email. Springer Nature journal authors may reuse their article's Version of Record, in whole or in part, in their own thesis without any additional permission required, provided the original publication is properly cited and includes the following acknowledgement "Reproduced with permission from Springer Nature". This includes the right to make a copy of your thesis available in your academic institution's repository, or other repository required by your awarding institution. For more information please visit see our FAQs [here](#).

If you have any further questions, please do not hesitate to get in touch.

Kind Regards,

Permissions Executive
SpringerNature
The Campus, 4 Crinan Street, London N1 9XW, United Kingdom
T
E
<http://www.nature.com>
<http://www.springernature.com>

Appendix B: Reference & copyright information by the publisher for the second essay (Essay II, Chapter 3)

Poszler, F., Geisslinger, M., Betz, J., & Lütge, C. (2023). Applying ethical theories to the decision-making of self-driving vehicles: A systematic review and integration of the literature. *Technology in Society*, 75, 102350. <https://doi.org/10.1016/j.techsoc.2023.102350>

Link to the Creative Commons user license: <https://creativecommons.org/licenses/by/4.0/>

(Formatting changes were made to align the published article with the format of this doctoral dissertation.)



Mi 07.02.2024 08:24

Permissions Helpdesk <permissionshelpdesk@elsevier.com>

Re: Permission letter for article inclusion in my doctoral dissertation [240206-042323]

An ☐ Franziska Poszler

 Zur Nachverfolgung. Erledigt am Mittwoch, 7. Februar 2024.

Dear Franziska Poszler,

Thank you so much for contacting us.

This article is available under the terms of the [Creative Commons Attribution License \(CC BY\)](#).

You may copy and distribute the article, create extracts, abstracts and new works from the article, alter and revise the article, text or data mine the article and otherwise reuse the article commercially (including reuse and/or resale of the article) without permission from Elsevier. You must give appropriate credit to the original work, together with a link to the formal publication through the relevant DOI and a link to the Creative Commons user license above. You must indicate if any changes are made but not in any way that suggests the licensor endorses you or your use of the work.

Permission is not required for this type of reuse, you can include this article as your chapter in your thesis.

All the best for your thesis submission!

Kind regards,



Senior Copyrights Specialist

ELSEVIER | HCM - Health Content Management

Visit [Elsevier Permissions](#)

Appendix C: Reference & copyright information by the publisher for the third essay (Essay III, Chapter 4)

Poszler, F., Geisslinger, M., & Lütge, C. (2024). Ethical Decision-Making for Self-Driving Vehicles: A Proposed Model & List of Value-Laden Terms that Warrant (Technical) Specification. *Science and Engineering Ethics*, 30(5), 47. <https://doi.org/10.1007/s11948-024-00513-0>

Link to the Creative Commons user license: <http://creativecommons.org/licenses/by/4.0/>


(Formatting changes were made to align the published article with the format of this doctoral dissertation.)




Fr 11.10.2024 14:09

Journalpermissions <journalpermissions@springernature.com>

RE: Submission Confirmation

An  Franziska Poszler

 Zur Nachverfolgung. Beginn am Freitag, 11. Oktober 2024. Fällig am Freitag, 11. Oktober 2024.

Dear Franziska,

Thank you for your recent Springer Nature permissions enquiry.

Your work available here: <https://link.springer.com/article/10.1007/s11948-024-00513-0> is licensed under the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, modification, and reproduction in any medium, provided you:

- 1) give appropriate acknowledgment to the original author(s) including the publication source,
- 2) provide a link to the Creative Commons license, and indicate if changes were made.

Therefore, you do not need a licence to reuse, republish, or adapt your material in your thesis.

Kind Regards,



Permissions Assistant
Springer Nature

The Campus, 4 Crinan Street, London N1 9XW, United Kingdom

T:

E:



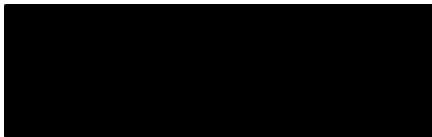
Appendix D: Author contributions to the three essays in this dissertation

Essay I: Formalizing Ethical Principles Within AI Systems: Experts' Opinions on Why (Not) And How to do it

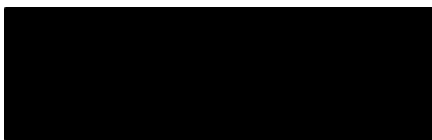
FP as the first author initiated the idea of this paper and contributed essentially to its conception and content. Material preparation, data collection and analysis was performed by FP and validated by EP and his team. FP wrote the first and final draft of the manuscript and EP and CL commented on previous versions of the manuscript. As corresponding author, FP was responsible for coordinating the submission and peer-review process.



Franziska Poszler (FP) (lead author)



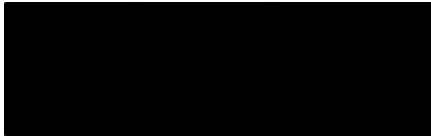
Edy Portmann (EP) (co-author)



Prof. Dr. Christoph Lütge (CL) (co-author)

Essay II: Applying Ethical Theories to the Decision-Making of Self-Driving Vehicles: A Systematic Review and Integration of the Literature

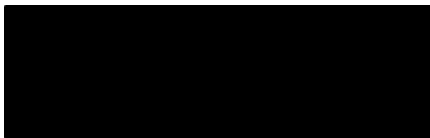
FP initiated the idea of this paper, conducted the literature search, and contributed essentially to its conception, design, and content. Data collection and analysis was performed by FP and validated by MG. FP wrote the first and final draft of the manuscript and MG, JB, CL critically reviewed and commented on previous versions of the manuscript. As corresponding author, FP was responsible for coordinating the submission and peer-review process.



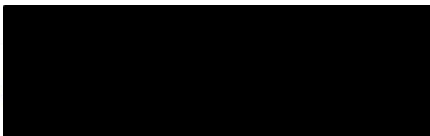
Franziska Poszler (FP) (lead author)



Maximilian Geisslinger (MG) (co-author)



Prof. Dr. Johannes Betz (JB) (co-author)



Prof. Dr. Christoph Lütge (CL) (co-author)

Essay III: Ethical Decision-Making for Self-Driving Vehicles: A Proposed Model & List of Value-Laden Terms That Warrant (Technical) Specification

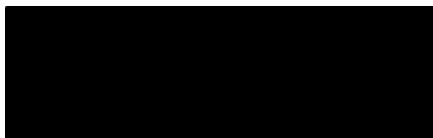
FP as the first author of this paper initiated the article's idea and contributed essentially to its conception and content. FP wrote the first and final draft of the manuscript and MG, CL critically reviewed and edited previous versions of the manuscript. As corresponding author, FP was responsible for coordinating the submission and peer-review process.



Franziska Poszler (FP) (lead author)



Maximilian Geisslinger (MG) (co-author)



Prof. Dr. Christoph Lütge (CL) (co-author)