Evangelische Theologische Faculteit, Leuven

Integrating Theological Ethics of Responsibility with Teleological and Deontological Ethics for Embodied AI

A Thesis Submitted in Partial Fulfillment of the Requirements for the degree of

Master of Arts in Theology and Religious Studies

in the Department of Systematic Theology

Advisor: Dr. Steven C. van den Heuvel

By
Michel Verhaegen

Leuven, Belgium
April 20, 2021

“*It is* the glory of God to conceal a thing: but the honour of kings *is* to search out a matter.”

Prov 25:2 - KJV

Abstract

This thesis addresses the ethics of responsibility in designing and using Artificial Intelligence (AI) in forthcoming technology. AI computer algorithms that – when embodied in technology products, such as a car – allow these products to conduct activities that were thus far exclusively performed by humans. Embodied AI processes real-life data, such as images of sceneries, that is imperfect due to, for example, bad lighting conditions or partially occluded images and, therefore, has to determine its actions based on imprecise data. The combination of the vast capabilities of embodied AI, which allows it to increasingly take over human activities, such as steering a car, and the imperfection of the AI software and the data that it uses present significant challenges to those responsible for designing and using this technology, as their decisions regarding this technology, may lead to life-threatening accidents. Its capabilities to take over the role of humans may cause so-called responsibility gaps whereby, loosely speaking, AI cannot take action due to the uncertainty in the data it perceives, while the human operator involved does not interfere based on the impression that when AI is active, no human interference is necessary. Such responsibility gaps are moral questions that can be studied in various normative ethical frameworks. Thus far, this has been done mainly from deontological and teleological traditions, from a secular perspective. In this thesis, I address these approaches in dialogue with theological ethics of responsibility, as developed in the work of the Protestant theologians Dietrich Bonhoeffer and Richard Niebuhr. Using the method of correlation proposed by the work of Paul Tillich, the contributions of Bonhoeffer on the structure of responsibility, and that of Niebuhr on the responsible self will be used to translate and apply these contributions to the ethics of responsibility in Automated Driving Systems (ADS) using AI. The dialogue between these theological frameworks to study and assess human responsibility and the existing secular

approaches results in four main contributions. The first two contributions address the design of AI for ADS and the last two are about design that enables the human operator to use ADS responsibly.

The first contribution from this dialogue brings forward and/or supports major criticisms about the classical ethical approaches to program ethics in AI. The major aspect of this critique is that these classical approaches only consider highly hypothetical traffic dilemmas in which all circumstances are precisely known. In this way, these approaches lack realism. Further, the outcome of these classical approaches to these hypothetical dilemmas is inconclusive and, as the discussion continues, they are not yet suitable for programming ethics in AI.

The second contribution frames the recent developments that aim to overcome the drawbacks of the classical approach analyzed in the first contribution by adopting the framework of Bonhoeffer's structure of responsibility. More precisely, it is shown that these developments, which provide the designer of AI for ADS with the tools to program ethically, are more in line with the structure of responsibility as formulated by Bonhoeffer, especially to act in line with reality (*Wirklichkeitsgemäßheit).* This framing illustrates the overall contribution I try to make in this thesis, namely that it encourages theologians and engineers to work together from the onset of the design of AI systems.

The third contribution is another critical analysis of a recent development in the ethics of responsibility of AI for ADS. By bringing the contributions of Bonhoeffer and Niebuhr in dialogue with a recent development to overcome the responsibility gap when using ADS, the fundamental starting point by which a human being is viewed between both frameworks of analyzing responsibility was highlighted. The recent development started from a reductionist view of human beings by assuming that decisions are made by a certain internal mechanism that needs to be identified and imitated by scientists. It resulted in making the human operator a ‘slave’ in ADS and may result in very inefficient driving when the human operator does not respond because car manufacturers have told drivers that when AI is active, they can enjoy a nap. Contrary to this reductionist view of human beings, Bonhoeffer and Niebuhr adopt a holistic view of human nature and its relationship to the Creator and His Son Jesus Christ. By this holistic view, the human is made a ‘master’ of the design and operation of ADS, anticipating that the system might have shortcomings or even demonstrates failures.

A ‘by-product’ of the critical analysis in the third contribution is the proposal of a new blueprint to make the human operator responsive and, hence, responsible even when ADS is doing most of the work. This new blueprint translates the four elements of the model of human responsibility that were postulated by Niebuhr in his book, *The Responsible Self,* into the ADS context. This translation of these four elements allows the user to gain improved insights into what is going on in the ADS while it is active and this involvement of the user will result in a mitigation of the responsibility gap while preserving efficiency. This comes at a price that calls the human operator to remain responsible all the time and will lead to an increased workload. I postulate that this ‘price’ will enable one to better achieve the anticipated improvements of ADS, such as increased safety, transport efficiency, and reduction of car pollution, compared to merely allowing AI to ‘blindly’ take over the role of the human operator. Finally, in addition to these improvements that result when one addresses responsibility from a theological perspective, theology is also helpful to offers insights to encourage human operators to put in this extra effort. This help consists of showing the effects that technology advancements, such as AI for ADS, have on the individualism of modern man combined with a Christological invitation to focus on the vulnerable ‘other’.

Acknowledgment

Thank you, God, for the opportunities You have and are given me to stand on the shoulders of scholarly giants who have gone before me.

Thank you, ETF professors and staff, for having the patience to teach me to critically study the Word of God and its many implications in human life, mine included.

Thank you, fellow ETF students with whom I shared classes, for showing me your perseverance and dedication and encouraging me to persist, despite busy daily schedules.

Thank you, Steven, for having the patience and courage to supervise an engineer who wants to take the first steps in becoming a theologian. Thanks for assisting me during these first exciting steps.

Thank you, Hilde, for being patient with me and for allowing me to continue studying, even after so many years already. Being a wife of a scientist is one thing, but of a soon-to-be theologian is perhaps something else altogether. I hope that it will be as exciting, or even more so!

Thank you, Dietrich Bonhoeffer and Richard Niebuhr, for leaving me with such a rich legacy. One lifetime is not enough to appropriate such richness.

Above all, I again, want to thank our Lord Jesus Christ for giving me a perspective to live for – a perspective where He will make all things new and where He is daring to invite us, human creatures, to participate. I commit this thesis into Your hands.

Contents

Chapter I

Introduction

This chapter first elaborates on the subjects ‘embodied Artificial Intelligence’ and the ‘ethics of responsibility’. After presenting the *status quaestionis* of the treatment of ethics for Artificial Intelligence (AI), I focus this thesis on the AI developments in Automated Driving Systems (ADS) and state my research questions for this class of AI applications. I then describe the methodology I use to answer these research questions and describe the outline of this thesis.

I.1 The Subject of Embodied AI

About half a century ago, Artificial Intelligence (AI) focused on developing and using computer programs whereby intelligence referred to the processing of pre-determined input (data) via fixed algorithms[1] to produce outputs in the form of commands or recommendations to the user. Widely cited successful examples are the data-mining programs of Google that, among other things, make recommendations on which video to watch next and the chess programs “that can beat 99.99 percent of all humans on earth.”[2] This computationally focused approach towards AI mainly attracted researchers “from computer science, psychology, philosophy, and linguistics.”[3]

[1] “An algorithm is a specific procedure for solving a well-defined computational problem. The development and analysis of algorithms is fundamental to all aspects of computer science: artificial intelligence, databases, graphics, networking, operating systems, security, and so on. Algorithm development is more than just programming. It requires an understanding of the alternatives available for solving a computational problem, including the hardware, networking, programming language, and performance constraints that accompany any particular solution. It also requires understanding what it means for an algorithm to be “correct” in the sense that it fully and efficiently solves the problem at hand.” (Encyclopedia Britannica, “Computer Science - Algorithms and Complexity,” accessed 23 March 2021, https://www.britannica.com/science/computer-science).

[2] Rolf Pfeifer and Fumiya Iida, “Embodied Artificial Intelligence: Trends and Challenges,” in *Embodied Artificial Intelligence: International Seminar, Dagstuhl Castle, Germany, July 7-11, 2003. Revised Papers*, ed. by Fumiya Iida et al., Lecture Notes in Computer Science (Berlin: Springer, 2004), 2, https://doi.org/10.1007/978-3-540-27833-7_1.

[3] Pfeifer and Iida, “Embodied Artificial Intelligence,” 5.

A distinction must be drawn between *computational* AI and *embodied* AI. In the early nineties of the previous century, Brooks[4] was one of the first to add the adjective 'embodied' to AI. Embodied AI systems are inherently embodied in the physical and social world. Consequently, they must address many issues that were not considered by computational AI. Instead of the precise inputs that form the basis of computational AI, embodied AI interacts in the real world with imprecise data such as that obtained through vision systems, which record partially occluded images of moving objects under changing lighting conditions. In such interaction, collaboration with humans also becomes a crucial and distinct aspect of embodied AI compared to computational AI. To deal with such interaction, the input-output behavior of the data processing is adapted. For example, when neural networks learn its input-output map to select a candidate (output) from records of their CVs (input) by only making use of male applicants, this input-output map will adapt when that neural network is subsequently offered female applicants in a new learning mode. When restricting to male applicants, the principle of "garbage in, garbage out" teaches that a (severe) bias in the selection will result.[5] This has led to the company Amazon, for example, suspending its recruiting engine when it was determined that it favored men over women.[6] Embodied AI has attracted researchers from computer science and

[4] Rodney A. Brooks, "Intelligence without Reason," in *Proceedings of the 12th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'91 (San Francisco, CA: Morgan Kaufmann Publishers, 1991), 569–95.

[5] R. Stuart Geiger et al., "Garbage In, Garbage Out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From?," in *Proceedings of the 2020 Conference on Fairness, Accountability and Transparency*, (Barcelona, December 17, 2019), 325–336, https://doi.org/10.1145/3351095.3372862.

[6] Jeffrey Dastin, "Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women," *Reuters*, October 10, 2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

philosophy, as with computational AI, but now also from “engineering, robotics, biology, and neuroscience (with a focus on dynamics).”[7]

Most of the current investment in the development of AI technology goes into machine learning.[8] Machine learning includes the field of deep neural networks, reinforcement learning, and meta-learning.[9] This survey paper of Sun, et al. explains that all (or most) machine learning algorithms are solving optimization problems.[10] In a so-called (un)supervised learning mode, neural networks can extract patterns from (un)labeled data. An example of the use of this learning feature is in *object recognition* for Automated Driving Systems. Generally, object recognition is conducted from an image containing multiple objects. For example, in a picture of a traffic scene where a pedestrian in front of your car is crossing a crosswalk secured by a traffic light, the aim of object recognition is threefold. First, it *localizes an object* in the picture. This is generally done by framing a single object (e.g., the pedestrian) with a (coarse) bounding box. Second, the localized object is then used in a subsequent *image classification* step. This step classifies or labels the type of object in the box based on the ‘resemblance’ of the actual object to the types in the neural network memory. As the actual object is generally not in the memory of the neural network and because the environmental and lighting conditions can differ greatly, this step can introduce uncertainty in the object recognition step. The final step is object detection,

[7] Pfeifer and Iida, “Embodied Artificial Intelligence,” 5–6.

[8] Renda, “Artificial Intelligence: Ethics, Governance and Policy Challenges,” *CEPS Task Force*, February 15, 2019, 14, https://www.ceps.eu/ceps-publications/artificial-intelligence-ethics-governance-and-policy-challenges/.

[9] For more details, see S. Sun et al., “A Survey of Optimization Methods from a Machine Learning Perspective,” *IEEE Transactions on Cybernetics* 50, no. 8 (August 2020): 3668–81, https://doi.org/10.1109/TCYB.2019.2950779.

[10] Sun et al., 3668.

which is a combination of the first two steps for each (relevant) object in the picture. For more details on recent developments in object recognition, refer to the work of Fujiyoshi, et al.[11]

AI is not an end in itself but is viewed "as a component in more complex information systems."[12] AI may be applied in many different areas. In medicine, it is used to improve computerized pathology of diseases or as robotic surgery assistants. In industry, it is used in advanced visual technology to enable robots to collaborate safely with humans. In economics, AI helps economists, governments, and others design tax policies to optimize the balance between social welfare and environmental protection. Lastly, in society, Automated Driving Systems (ADS) claim to take over all functions of the human driver. Such systems will significantly influence our societies. Apart from the many concrete AI developments that are on the drawing board and that are appearing or will soon appear in our societies, some authors take a quantum leap forward in their prediction of the impact of AI on future technology developments. One such author is the historian Yuval Noah Harari, who, in his book *Homo Deus: A Brief History of Tomorrow*,[13] predicts that humans may initially develop a seed algorithm that will result in "the use of machine learning and artificial neural networks [evolving] independently ... and [following] their own path, going where no human has gone before — and where no human can

[11] Hironobu Fujiyoshi, Tsubasa Hirakawa, and Takayoshi Yamashita, "Deep Learning-Based Image Recognition for Autonomous Driving," *IATSS Research* 43, no. 4 (December 1, 2019): 245, https://doi.org/10.1016/j.iatssr.2019.11.008.

[12] Renda, "Artificial Intelligence: Ethics, Governance and Policy Challenges," 11.

[13] Yuval Noah Harari, *Homo Deus: A Brief History of Tomorrow*, Illustrated edition (New York, NY: Harper & Row, 2017).

follow."[14] These super-algorithms, which are called superhumans, will eventually "make homo sapiens an obsolete algorithm."[15]

In this thesis, I focus on the embodied AI development in use or that will soon appear on the market, as in the fields of medicine, industry, economics, or society, as listed above. Such developments, on the one hand, are promoted based on the expected increased safety and complementarity with human abilities, which are often imprecise, sluggish, and demonstrate limited reliability. On the other hand, as these machines systematically take over decisions and responsibilities from humans, they may strip humans of their freedom. This ambiguity is present both in the design and in the use of these automated systems. In the case of design, ADS designers must make decisions on how to handle life-threatening car accidents beforehand, and users of automated driving cars may face the personal dilemma of whether or not to supervise this technology during its operation to override it when it malfunctions.

I.2 Ethics of Responsibility

When these embodied AI systems make decisions that possibly endanger human lives and also when they are using, possibly together with other automated systems and/or humans, restricted spaces in which conflicts with other 'users' may occur, we touch upon ethical questions.

This is illustrated in the following example.

Example 1: In May 2018, one of Uber's self-driving test cars crashed into and killed a jaywalking pedestrian. The bad illumination of the road in front of the Uber vehicle caused the

[14] Harari, *Homo Deus: A Brief History of Tomorrow*, 458.

[15] Harari, 444.

object recognition in the AI software of the self-driving car to frequently change its classification of the object in front of the vehicle from vehicle to other to bicycle in the last 5.6 seconds before the crash.[16] This indetermination contributed to the uncertainty that caused ADS not to intervene. Moreover, as the ADS user also did not intervene, a fatal crash resulted. This example raises various moral questions for the designers and users of ADS. For example, one could ask the designers of the AI software, "What should be the right or better thing to do in dealing with the uncertainty in the AI object recognition?" Furthermore, one could ask the ADS user, "How could he have contributed to avoiding or mitigating such accidents?"

The moral questions highlighted in Example 1 represent aspects of the general definition of responsibility used by the Centre for European Policy Studies (CEPS) task force in their report on *Artificial Intelligence: Ethics, Governance, and Policy Challenges*.[17] In this case, responsibility in the context of AI is stated as follows: "Responsibility implies acknowledging the potential risks of AI, and accordingly acting to mitigate them in the design, development, and use of AI."[18]

Example 1 illustrates a relevant responsibility problem concerning embodied AI, namely the responsibility gap. This is, loosely stated, the vacuum that arises in the 5.6 seconds before the crash, resulting in neither the ADS nor the human operator intervening in this life-threatening traffic scenario. The vacuum resulted from the confluence of, on the one hand, the uncertainty in the data processed by the ADS AI, causing it to internally change its course of action so

[16] Katyanna Quach, "Remember the Uber Self-Driving Car That Killed a Woman Crossing the Street? The AI Had No Clue about Jaywalkers," *The Register*, November 6, 2019, https://www.theregister.com/2019/11/06/uber_self_driving_car_death/.

[17] Renda, "Artificial Intelligence: Ethics, Governance and Policy Challenges."

[18] Renda, 30.

frequently that inaction resulted, and, on the other hand, the human operator remaining inactive as he was convinced ADS was doing (a safe) job. The notion of the responsibility gap was first introduced by Andreas Matthias concerning possible gaps in responsibility when different humans and automated machines are working together. The definition is as follows:

> Presently, there are machines in development or already in use, which are able to decide on a course of action and to act without human intervention. The rules by which they act are not fixed during the production process but can be changed during the operation of the machine, by the machine itself. This is what we call machine learning. Traditionally, we hold either the operator/manufacture of the machine responsible for the consequences of its operation or "nobody" (in cases where no personal fault can be identified). Now it can be shown that there is an increasing class of machine actions where the traditional ways of responsibility ascription are not compatible with our sense of justice and the moral framework of society because nobody has enough control over the machine's actions to be able to assume responsibility for them.[19]

Based on replacing robots with machines in the shortened definition of Sven Nyholm, I arrive at the following, more compact, definition, namely, the responsibility gap occurs when *an autonomous machine is acting outside the control or oversight of a human agent.*[20] This thesis will use this more compact definition of the responsibility gap.

Ethics does not coincide with moral questions, but it is "the field of study, or branch of inquiry, that has morality as its subject."[21] Combining the CEPS definition of responsibility with the latter definition of ethics, the ethics of responsibility for AI can now be formulated as follows:

> Ethics of Responsibility for AI is ethical research or investigation that systematically addresses moral questions under the guiding principle of responsibility in the context of AI.

[19] Andreas Matthias, "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata," *Ethics and Information Technology* 6, no. 3 (September 1, 2004): 177, https://doi.org/10.1007/s10676-004-3422-1.

[20] This definition is based on that used in Sven Nyholm, "Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci," *Science and Engineering Ethics* 24, no. 4 (August 1, 2018): 1213, https://doi.org/10.1007/s11948-017-9943-x, where machines would then be only restricted to robots.

[21] Wolfgang Huber, "Ethics of Responsibility in a Theological Perspective," *Stellenbosch Theological Journal* 6, no. 1 (2020): 187, https://doi.org/10.17570/stj.2020.v6n1.a11.

Such a systematic approach can be pursued from a secular, anthropocentric perspective by the application of ethical frameworks of deontology or utilitarian ethics, but it can also be pursued from a theocentric perspective. This thesis aims to bring the secular approach in dialogue with a theocentrically based ethics of responsibility for AI. This is based on the insight formulated also by Wolfgang Huber that ethics of responsibility has both an anthropological basis as well as a theological basis. The anthropological basis sees humans as "communicative beings, as people, listening to their call and answering it."[22] The theological basis is in "the dialectic of commitment and freedom."[23] It is about this theological dimension of commitment and freedom that Dietrich Bonhoeffer has made a firm contribution as they together define his "structure of responsible life." This is elaborated in more detail in Section IV.3.

I.3 *Status Quaestionis*

The ethics of responsibility of AI has been primarily investigated from a non-theistic perspective. Theologians have investigated the topic of ethics and AI; see, for example, the recent article of Alexis Fritz, et al.[24] This paper concludes that the evaluated existing models that intend to fully integrate AI into their concept of 'moral agency' are inadequate and recommends distinguishing "conceptually between the different entities, causalities, and relationships in a human-computer interaction, arguing that this is the only way to do justice to both human responsibility and the

[22] Huber, "Ethics of Responsibility in a Theological Perspective," 196.

[23] Huber, 196.

[24] Alexis Fritz et al., "Moral Agency without Responsibility? Analysis of Three Ethical Models of Human-Computer Interaction in Times of Artificial Intelligence (AI)," *De Ethica* 6, no. 1 (June 30, 2020): 3–22, https://doi.org/10.3384/de-ethica.2001-8819.20613.

moral significance and causality of computational behavior." How that distinction should be made and how humans should responsibly deal with AI is left open by that paper.

For the ADS development, according to the knowledge of the author of this thesis, the non-theistic perspective is the sole perspective taken.

Such secular investigations have been conducted by individual researchers or small teams of researchers to address a particular aspect of this research topic. The book of Lin, et al, *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, is an example of this.[25] In this collection, for example, the topic of how a designer of the ethics for AI selects between deontological or utilitarian ethics is addressed by Bhargava, et al.[26] This paper also addresses uncertainty, but, in this case, the uncertainty is the consequence and actualization of the multiple choices in normative ethical frameworks that the designer of the ethics for AI faces. Such uncertainty is different from the internal uncertainty in the AI processing due to the uncertainty in the data it is processing, as illustrated in Example 1. The contributions in this book and many others that are reviewed in Chapter III focus on hypothetical traffic scenarios that might not occur in reality. In addition to this lack of realism, which this thesis attempts to overcome, they do not consider the important facet of responsibility *during* the operation of AI, as was considered by the CEPS definition of responsibility for AI.

In addition to individual researchers, agencies representing bodies of researchers, or even continents, have considered the topic of ethics for AI. They prioritized the topic "to ensure that

[25] Patrick Lin, Keith Abney, and Ryan Jenkins, eds., *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence* (Oxford, New York, NY: Oxford University Press, 2017), https://doi.org/10.1093/oso/97880190652951.001.0001.

[26] Vikram Bhargava and Tae Wan Kim, "Autonomous Vehicles and Moral Uncertainty," in *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, edited by Patrick Lin, Keith Abney, and Ryan Jenkins (Oxford, New York, NY: Oxford University Press, 2017), 5–19, https://doi.org/10.1093/oso/97880190652951.001.0001.

the positive disruption and empowerment effects of AI prevail over the potential negative effects."[27] This concern stimulated a global dialogue on AI and resulted in various codes of ethics and declarations about AI. Examples include the "Asilomar principles," the "Declaration of Toronto," "AI for Good," but also corporate ethical principles about AI as issued by companies such as Google, SAP, IBM, Microsoft, Deutsche Telecom, and Telefonica.[28] The EU Centre for European Policy Studies (CEPS) synthesizes these contributions. They nominated a task force of forty experts in ethics and AI to help AI developers, organizations, and companies develop and use AI to balance the benefits of AI in the EU with responsibility, thus using the definition of responsibility as stated on page 17. This CEPS task force formulated three possible ways to operate responsibly as well as to increase trust in AI.

The first method uses and develops AI in *complementarity* with human beings. Such complementarity should lead to a win-win situation that combines the excellence of human beings with those of AI. According to the CEPS task force, humans are better than AI at "setting goals, using common sense, and formulating value judgments," while AI "may be better at pattern discovery, large-scale math, and performing statistical reasoning."[29] The second method deals with *bias and value alignment*. Making AI researchers aware of the bias problem, as illustrated above by Amazon's recruiting example, will stimulate research activities to make AI more resilient to such a bias by improved selection and filtering of the training data.[30] Value alignment refers to the criteria humans find acceptable in ethical decision-making, such as "life

[27] Renda, "Artificial Intelligence: Ethics, Governance and Policy Challenges," 5.

[28] Renda, 5–6.

[29] Renda, "Artificial Intelligence: Ethics, Governance and Policy Challenges," 27.

[30] See, for example, the work Geiger et al., "Garbage In, Garbage Out?" 335.

or death decisions."[31] In this case, the challenge is determining which ethical approach AI should use to make ethical decisions. I address this question in more detail concerning ADS in Chapter III. The third method relates to sustainability and encompasses social sustainability, addressing issues such as hunger, social inequality, and poverty, as well as environmental sustainability, including AI's carbon footprint. These three general routes for addressing the risks of AI technology are still rather vague. They are suggestive and are based on good intentions but they do not provide strategies for dealing with concrete ethical questions and challenges. A more systematic approach taken by the EU to make AI trustworthy is based on human rights. In a team of 52 experts who operated under the label High-Level Expert Group on AI (AI HLEG), such an approach led to the formulation of principles and recommendations as summarized in the report *Ethics Guidelines for Trustworthy AI*.[32] Grounded in the fundamental human rights of dignity, freedom, equality, and solidarity as enshrined in the EU Treaties, the EU Charter, and international human rights laws, the following four principles were formulated, specifically concerning AI:[33]

1. *Respect for human autonomy*: AI systems should not limit the self-determination of human beings, as the making of meaningful choices should be left to them. This means securing human oversight over work processes in AI systems. AI should support human beings, and AI should aim for the creation of meaningful work.

[31] Renda, "Artificial Intelligence: Ethics, Governance and Policy Challenges," 32.

[32] Nathalie Smuha, "Ethics Guidelines for Trustworthy AI," *European Commission*, April 8, 2019, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[33] Smuha, 12.

2. *Prevention of harm*: AI should not cause mental or physical harm to human beings in an environment where AI systems are active. This includes avoiding the "adverse impacts due to asymmetry of power or information, such as between employees and employer," and preventing harm to the environment and other living beings.[34]
3. *Fairness*: AI should help stimulate the equal and just distribution of benefits and costs, as well as the avoidance of unfair bias towards (groups of) individuals. AI developers should foster equal opportunity for everyone to AI education, its goods, services, and technology. The avoidance of unfair bias also means that the entity that makes the decision must be identifiable, and the decision process of the AI system should be explicable (as explained in the next principle).
4. *Explicability*: An AI component in a system is often considered a "black box." In response to this, to make the operation of AI transparent, dedicated measures need to be taken. For example, measures referred to in requirement four of the seven listed below. The degree of explicability depends on the degree of accuracy of the AI output. That means that when the output of AI is more certain, less explication is needed. For instance, in Example 1, when the object recognition recognizes an object that is correctly classified, no further explication is required. However, when it (frequently) makes incorrect classifications, explanations, and further analysis are required.

[34] Smuha, 12.

These principles offer a framework that may act as a general guide for AI designers and users to strive to adhere to. However, special fine-tuning is necessary because these principles may be conflicting in certain AI applications. For example, in ‘predictive policing’, the effort to reduce crime may create conflict between the principle to prevent harm and the principle of individual liberty and privacy. Such conflict arises when using the tool of predictive policing as a form of social control. The deliberation of these principles seeking “trade-offs via reasoned, evidence-based reflection rather than intuition” is still necessary,[35] taking into consideration the particular AI application and context. Another reason for further fine-tuning is that these principles are still rather abstract. Therefore, AI-HLEG has compiled the following list of seven requirements that AI practitioners should respect:

1. *Human Agency and Oversight*: This requirement stipulates that AI systems should respect the principle of human autonomy. Whenever there appears to be the risk that the AI system will negatively affect fundamental human rights, a full impact assessment must take place before the system’s development. This should include an evaluation of whether those risks can be reduced or justified as necessary in a democratic society to respect the rights and freedoms of others. The right of individual autonomy should be preserved by not subjecting the individual to decisions solely based on automated processing that produces legal effects on users or that affect them significantly. Human oversight should enable a human-in-command approach, whereby a human being can decide when, and how, to use the system in any particular situation. The general rule should apply that “the less oversight a

[35] Smuha, 13.

human can exercise over an AI system, the more extensive testing and stricter governance is required."[36]

2. *Technical Robustness and Safety*: This is a crucial component to achieve trustworthy AI. The requirement is expressed in utilitarian terms as "minimizing unintentional and unexpected harm, and preventing unacceptable harm."[37] It requires security and protection against malicious actors, such as hackers, for example. Moreover, using the example of ADS systems, from a vehicle safety perspective, a preventive approach is needed to cope with the unpredictability of AI in ADS systems. The recommendation given here is to provide a backup plan either on the level of the AI software, for example, going from probabilistic reinforcement learning control to rule-based procedures, or reverting to an external operator to continue the action. A final key element for this requirement is the validation and certification of AI-equipped systems. Reproducible tests, evaluating the proper operation of the AI-equipped system for a range of inputs and in a range of situations, contribute to its reliability.
3. *Privacy and Data Governance*: This requirement is closely linked to the principle of prevention of harm. It requires that the AI system in question protects user data and privacy, and the prevention of training AI systems with malicious data as well as organizations using the AI data should have clear protocols of who can access the data and under which circumstances.

[36] Smuha, 16.

[37] Smuha, 16.

4. *Transparency*: This requirement entails that data sets, actions, and environmental conditions are traceable through proper documentation, data labeling, and identification of the reasons why an AI decision was erroneous, as well as making AI systems understandable and traceable for human users and experts.
5. *Diversity, Non-discrimination, and Fairness*: This is closely linked to the principle of fairness. It requires equal access and equal treatment by avoiding unfair bias and making AI products and services accessible to "all people regardless of their age, gender, abilities, or characteristics."[38] Further, it is advisable to solicit regular feedback even after the deployment of stakeholders.
6. *Societal and Environmental Well-being*: Developers of AI systems should be encouraged to develop the system in a sustainable and ecologically responsible manner. As AI systems may enhance or diminish people's social skills, these effects need to be carefully monitored. This monitoring should include the impact that AI systems may have on democratic processes, such as in political decision-making or protecting the electoral process from misusing data of voters or misinforming them.
7. *Accountability*: To ensure responsibility and accountability for AI systems and their outcomes, the algorithms, data, and design processes should be auditable, taking into account the protection of intellectual property. To help minimize the potentially negative impact of AI systems, especially to those (in)directly affected, "red

[38] Smuha, 18.

> teaming"[39] or forms of "Algorithmic Impact Assessment"[40] before and during the development, deployment, and use of AI systems are recommended. If things go wrong, adequate action for redress should be foreseen and ensured.

The above recommendations can be used to compile a code of conduct for AI practitioners. However, such a "code of conduct might produce a 'tick box' culture of ethical complacency where complying with paperwork becomes an end in itself and the goal of ethical compliance is focused on too narrowly, and for the sake of reward or avoiding penalties."[41] An alternative to avoid being dragged in such a paper swamp, which was also recommended in the AI-HLEG report, was to tailor ethical investigations to the specific AI case studies.[42] Based on this overview of the ethics in AI, in the next section, I state the research question.

I.4 The Research Question

Following the recommendation of the AI-HLEG study to tailor the ethics of responsibility to a specific case study, this thesis focuses on ADS. Furthermore, I will bring theology into dialogue with the secular approaches to analyze and develop the ethics of responsibility for ADS. Therefore, I have formulated the following central research question:

[39] "Red teaming is the practice whereby a "red team" or independent group challenges an organisation to improve its effectiveness by assuming an adversarial role or point of view. It is particularly used to help identify and address potential security vulnerabilities. Smuha, "Ethics Guidelines for Trustworthy AI," 37.

[40] "Algorithmic Impact Assessment as a tool is a scorecard intended to bring attention to design and deployment decisions that might have been overlooked." Mathieu Lemay, "Understanding Canada's Algorithmic Impact Assessment Tool," *Medium*, 11 June 2019, https://towardsdatascience.com/understanding-canadas-algorithmic-impact-assessment-tool-cd0d3c8cafab.

[41] Paula Boddington, *Towards a Code of Ethics for Artificial Intelligence*, Artificial Intelligence: Foundations, Theory, and Algorithms (New York, NY: Springer International Publishing, 2017), 54, https://doi.org/10.1007/978-3-319-60648-4.

[42] Smuha, "Ethics Guidelines for Trustworthy AI," 3.

> *How can the theological ethics of responsibility, especially as developed by Dietrich Bonhoeffer and Richard Niebuhr, be integrated with already existing teleological and deontological frameworks of ethics, for the evaluation of the use and development of embodied AI, specifically related to Automated Driving Systems?*

In ADS, the tasks of the driver, such as lane-keeping and overtaking, are taken over by an AI system. Such systems, for example, automatically recognize objects in front of the vehicle and adapt the course of the vehicle to proceed safely. For ADS, the above-stipulated research question evokes two dedicated, more detailed sub-questions. First, "How could we integrate ethical decision-making into handling life-threatening traffic dilemmas as well as in mundane traffic operations?" Second, "When considering human-AI interaction, how should one enhance, or make possible, that humans can take responsibility even when the ADS does most of the work?" The latter question is stimulated by the human-centric views of both the EU CEPS task force and the approach of AI-HLEG, which view developments of AI as still in complementarity with human beings. However, in that complementarity, concerning the question of responsibility, responsibility gaps still arise.

I.5 Methodology

Bringing theology in dialogue with ethical frameworks to address the ethics of responsibility for ADS from a secular perspective is challenging for at least two reasons. First, you could ask 'What has ADS to do with theology?' When postulated in this cold fashion, one could answer 'nothing'. However, if one focuses on responsibility, more can be said. Therefore, I introduce the voices of Bonhoeffer and Niebuhr. Both renowned theologians have devoted a substantial part of their ethics to responsibility. For Bonhoeffer, this is, for example, documented in the section

"Structure of Responsible Life" in his *Ethics*-manuscript "History and Good [2],"[43] while for Niebuhr we have the compiled book *The Responsible Self*.[44]

Though both wrote extensively on the ethics of responsibility, they wrote in a different time and context. Furthermore, they did not explicitly write about AI, let alone about ADS. To make use of their profound insights about the topic of responsibility, a hermeneutical process to translate between Bonhoeffer's and Niebuhr's historical context and that of our own, as well as between their theology and contemporary developments in responsible ethics for ADS is necessary. This hermeneutical process constitutes the second challenge. To make such translation possible, use will be made of the method of correlation.

In systematic theology, the method of correlation came to be known through its description and use by Paul Tillich. In the first volume of his *Systematic Theology*,[45] he argues that correlation makes an analysis of the human situation out of which the existential questions arise, and it demonstrates that the Christian message can provide the answers to these questions. In explicating and using this method, Tillich's overriding intention "was apologetic, namely, to endow theology with greater relevance in an increasingly secular world."[46] Since the work of Tillich, the method has been adapted and criticized. Of the key criticism of the methods as presented by Francis Schüssler Fiorenza, the following three can be considered as major:[47]

[43] Dietrich Bonhoeffer, *Ethics*, ed. by Clifford J. Green, trans. by Reinhard Krauss, Charles C. West, and Douglas W. Stott, Dietrich Bonhoeffer Works, vol. 6 (Minneapolis, MN: Fortress Press, 2005), 246–298.

[44] H. Richard Niebuhr, *The Responsible Self: An Essay in Christian Moral Philosophy* (New York, NY: Harper & Row, 1963).

[45] Paul Tillich, "The Method of Correlation," In ibid., Systematic Theology, vol. 1 (Chicago, IL: University of Chicago Press, 1950), 59–66.

[46] Steven C. van den Heuvel, *Bonhoeffer's Christocentric Theology and Fundamental Debates in Environmental Ethics* (Eugene, ORE: Pickwick Publications, an Imprint of Wipf and Stock Publishers, 2017), 13.

[47] Heuvel, *Bonhoeffer's Christocentric Theology and Fundamental Debates in Environmental Ethics*, 14.

1. The erroneous presupposition that the language used to describe reality equals reality itself.
2. The inadequate consideration of change and non-identity in the development of faith and theology.
3. The failure to adequately criticize the Christian tradition, for, while it may take issue with certain theological formulations, it does not criticize the underlying 'experiences and affirmations' that these formulations express."[48]

To address the adaptations of the method of correlations and the above criticism, Steven C. van den Heuvel developed a variant of the correlation method; he does so by translating the ethics of Bonhoeffer to the actual problem of environmental ethics in *Bonhoeffer's Christocentric Theology and Fundamental Debates in Environmental Ethics.*[49] The three constitutive elements of this variant will also be used for the hermeneutical process of translation in this thesis. These elements are the following: 1) the interpretation of Bonhoeffer's and Niebuhr's work, 2) my interpretation of several contributions in the ethics of responsibility for AI, and 3) the perspective from which the first two elements are brought into correlation. These three elements are described briefly:

1. Principles for interpreting Bonhoeffer's and Niebuhr's work on the ethics of responsibility. In this case, the focus is on these concepts in the work of Bonhoeffer and Niebuhr that can contribute to the understanding of recent developments in the ethics of responsibility of AI for ADS, as well as on mitigating existing responsibility gaps. This focus will be achieved by (a) discovering overlapping concepts in the

[48] Heuvel, *Bonhoeffer's Christocentric Theology and Fundamental Debates in Environmental Ethics*, 14.

[49] Heuvel.

works of Bonhoeffer and of Niebuhr, which are relevant for the analysis of contemporary ethics of responsibility for AI in ADS, (b) establishing a critical dialogue with other scholars about the work of Bonhoeffer and Niebuhr, and (c) reviewing the work of other scholars and their use of correlation to analyze (possibly) other contemporary ethical questions.

2. Addressing the contemporary field of the ethics of responsibility for ADS. Most ethical frameworks that analyze and design the ethics of responsibility are deontological or teleological. A review will examine how these two frameworks have been used to address the ethics of responsibility.
3. Correlation between Bonhoeffer's and Niebuhr's work on the ethics of responsibility to contemporary secular accounts for ADS. To bring these contemporary ethics as applied to ADS in dialogue with the work of Bonhoeffer and Niebuhr, I first compare the theological ethical frameworks of Bonhoeffer and Niebuhr with each other. These common insights will then be used to understand recent improvements, to criticize them, and possibly improve them. The latter especially focuses on the mitigation of the responsibility gap.

I.6 Overview of the Thesis

The outline of the thesis is as follows. In Chapter II, the ADS case study is described in more detail. Based on the taxonomy of different levels of ADS, the level of ADS that is currently widely tested by different car manufacturers and for which real-life data about ethical issues in handling accidents with such ADS have been reported will be taken as a central case throughout this thesis. Chapter III summarizes the literature on the application of teleological and deontological ethics to address the first sub-question of this thesis. The shortcomings of the presented approaches motivate further study of the alternative ethical frameworks of Bonhoeffer and Niebuhr. I will describe the structure of responsible life, according to Bonhoeffer, and the responsible self, according to Niebuhr, in Chapter IV, under the umbrella of "Theological Ethics of Responsibility." This thesis will demonstrate that while these views were compiled at least half a century ago, they are still very useful for addressing contemporary modern ethical problems. A first application, mainly for increased realism by Bonhoeffer's structure of responsible life, is presented in Chapter V. Here I will address the first sub-question of this thesis.

To address the second sub-question of this thesis, Chapter VI introduces the recent method of meaningful human control over ADS by Santonio and van den Hoven.[50] This method presents a solution to the responsibility gaps in ADS where humans should act as back-ups in case of ADS failure. The method of meaningful human control over ADS is reviewed in Chapter VII using Niebuhr's analysis of teleological ethics. In that same chapter, Niebuhr's work is further used to formulate the theoretical ground for an alternative new solution to address the

[50] Filippo Santoni de Sio and Jeroen van den Hoven, "Meaningful Human Control over Autonomous Systems: A Philosophical Account," *Frontiers in Robotics and AI* 5 (February 18, 2018), https://doi.org/10.3389/frobt.2018.00015.

responsibility gap. Chapter VIII then presents the concretization of this alternative new solution. Finally, in Chapter IX, I review the answers to the research question and its two derived sub-questions, and I analyze the possible contributions these answers can provide to the field of ethics for ADS, in particular, and autonomous systems, in general. Moreover, I challenge Christian theologians with expertise in ethics and an interest in equipping society to responsibly deal with the surge of technological innovation that will substantially overwhelm (Western) society in the coming years. One such challenge is to break the trend of individualism, which views new technology as a status symbol rather than a tool. To break this trend and challenge people to care for the vulnerable other, a radical change in the pattern of thinking of the modern man is required. Furthermore, as Niebuhr has said regarding making such change, our sense of the ultimate context needs to be revised. This is where Christian theology can make a difference.

Chapter II

Automated Driving Systems

II.1 Introduction

Developments in AI and control engineering have made it possible for several key activities of the human driver to be superseded by automated subsystems. Examples are "automatic braking to maintain lane position in traffic or to avoid a sudden obstacle or hazardous event in the vehicle's pathway" or automatic lane changing.[51] Many important players in the car industry and AI companies are early adaptors of this new technology and have shown a great interest in these developments, including Google, Aurora Tech, Tesla, Volvo, Mercedes, Volkswagen, and Toyota. These companies 'sell' this new technology under various nametags, such as "feature complete self-driving cars,"[52] "unsupervised driving,"[53] self-driving, (semi-) autonomous driving, and unmanned vehicles. This technology is being promoted as having major potential to improve driving safety, reduce emission, make 'easy' transportation available to mobility-impaired people, and increase efficiency. The annual benefit to society by deploying this technology on a wide scale is projected to reach almost $800 billion by 2050 "through

[51] SAE, "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles," *Technical*, June 15, 2018, 34, https://www.sae.org/standards/content/j3016_201806/.

[52] Lance Eliot mentioned in his Forbes article the announcement of Tesla's latest development in its self-driving car program in October 2019 that Elon Musk referred to "complete self-driving cars". Eliot tried in a communication with Elon Musk to clarify this term, but without success. Therefore, Eliot went out on a limb, suggesting the meaning to be "to able to be autonomous but requiring supervision and intervention at times." See Lance Eliot, "Has Elon Musk Set Up Regulatory Boogeyman As Scapegoat For Ongoing Delay In Promise Of Self-Driving Teslas?," *Forbes*, April 17, 2020, https://www.forbes.com/sites/lanceeliot/2020/04/17/has-elon-musk-set-up-a-regulatory-boogeyman-as-scapegoat-for-delay-in-his-self-driving-tesla-promise/.

[53] This terminology is used by Volvo cars. See Volvo Cars, "Autonomous Driving | Intellisafe | Volvo Cars," accessed June 26, 2020, https://www.volvocars.com/en-kw/own/own-and-enjoy/autonomous-driving.

congestion mitigation, road casualty reduction, decreased energy consumption, and increased productivity caused by the reallocation of driving time."[54]

To investigate self-driving cars, it is necessary to clarify the taxonomy of terminology of different degrees or levels in automated driving. Therefore, Section II.2 describes the taxonomy developed by the Society of Automotive Engineers (SAE). Section II.3 then highlights the type of automation that will be mainly considered as an object for the ethical analysis of this thesis.

II.2 SAE Taxonomy of Driving Automation Systems

The taxonomy of SAE is based on Michon's classification of the acts of driving into three types of driver functional levels: "Strategic, Tactical, and Operational."[55] The strategic level involves trajectory planning, including the determination of the waypoints, the destination, the best routes to take, and the desired time of arrival. The tactical level involves maneuvering the car in the actual traffic. At this level, the car is maintained at the appropriate speed, lane changes are performed, and decisions about whether or not to overtake are made. At the operational level, vehicle control manipulations are performed involving split-second reactions "that can be considered pre-cognitive or innate,"[56] such as braking, accelerating, keeping lane position in traffic, and avoiding obstacles or a hazardous event in the car's trajectory. These three functional levels operate in a coordinated hierarchy whereby the top level is the strategic level. This level

[54] Ekim Yurtsever et al., "A Survey of Autonomous Driving: Common Practices and Emerging Technologies," *IEEE Access* 8 (2020): 1, https://doi.org/10.1109/ACCESS.2020.2983149.

[55] John A. Michon, "A Critical View of Driver Behavior Models: What Do We Know, What Should We Do?" in *Human Behavior and Traffic Safety*, ed. by Leonard Evans and Richard C. Schwing (Boston, MA: Springer US, 1985), 485–524, https://doi.org/10.1007/978-1-4613-2173-6_19.

[56] SAE, "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles," 34.

coordinates and constrains the lower-level tasks of tactical maneuvering. The latter level coordinates and constrains the lowest level tasks of operational control.

Based on these three functional levels of driving acts, two important notions need to be introduced before introducing the six levels of Automated Driving Systems. Firstly, the Dynamic Driving Task (DDT) in the SAE taxonomy comprises the levels of tactical maneuvering and operational control. These two levels are meant to operate the vehicle in real-time while considering the on-road traffic. Secondly, the six detailed operation/tactical driving tasks are derived as follows:[57]

1. Lateral vehicle motion control by steering the vehicle at the operational level.
2. Longitudinal vehicle motion control by acceleration and braking – also at the operational level.
3. Monitoring the environment of the car via object[58] recognition and response preparation. These activities are a mixture of operational and tactical tasks.
4. Object response execution at the tactical level.
5. Manoeuver planning at the tactical level, and
6. Enhancing the visibility of the car by switching on the headlights, using the direction indicators, fog lamp, etc. This is again comprised of activities at the tactical level.

With this organization of the driving task, different levels of automation can now be defined. Following the SAE standards, the following six levels of automation are given:[59]

[57] SAE, 6.

[58] An object in traffic includes both static objects, such as stop lights, traffic signs, but also moving objects, such as neighboring cars, bicycles, mopeds, pedestrians, etc. The tasks in Object Recognition have been briefly defined on page 7.

[59] SAE, "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles," 24–25.

1. Level 0 — *No driving automation:* The driver performs the full DDT, possibly assisted by *active safety systems*, such as electronic stability control and automated emergency braking, and certain types of driver assistance systems, such as lane-keeping assistance.
2. Level 1 — *Driver assistance:* The driving automation system performs a sustained and Operational Design Domain (ODD) specific, for example, on highways, driving either in the lateral vehicle motion control mode or the longitudinal vehicle motion control mode, but not both simultaneously. Thus, only limited Object Detection and Response (ODR) tasks are performed by the automated systems and, therefore, the driver must supervise the driving automation system's performance by completing all remaining activities of a full DDT.
3. ADS level 2 — *Partial driving automation*: The driving automation system performs sustained and ODD-specific driving for both the lateral and longitudinal vehicle motion control subtasks of DDT. The driver is expected to complete the ODR subtasks and *supervise* the driving automation system. The latter supervision is needed as the ODR cannot recognize or respond to some objects.
4. ADS level 3 — *Conditional driving automation*: The driving automation system performs sustained and ODD-specific driving for the entire DDT with the exception that the DDT can revert to a ready user when the ADS issues a request to intervene, as well as when system failures occur in other vehicle systems, who will then respond appropriately. The "DDT fall-back ready user need not supervise an ADS level 3 while it is engaged but is expected to be prepared to either resume the DDT when the ADS issues a request to intervene or to perform the fall-back and achieve a minimal

risk condition if the failure condition precludes normal operation."[60] The minimal risk condition means that a typical driver is provided sufficient time to respond appropriately to the driving situation at hand. The backup user is also expected to be receptive to vehicle failures, even when ADS does not trigger a request to intervene.

5. ADS level 4 — *High driving automation*: The driving automation system performs sustained and ODD-specific driving for the entire DDT, including DDT backup without any exception that calls on the user to intervene. The user, hence, does not need to supervise the ADS or be receptive to intervene upon request, provided the ADS is engaged. The ODD is restricted, such as in closed campus shuttles or a high-speed cruising featured highway. Outside these restricted areas, the driver must take over the driving.
6. ADS level 5 — *Full Driving Automation*: The driving automation system performs sustained and unconditional (no specific ODD requirements) driving for the entire DDT including DDT fallback without any exception that calls on the user to intervene. The ADS automatically performs DDT fallback in case of performance-relevant system failures of the ADS or vehicle. The user does not need to be receptive or to intervene if the ADS is engaged.

It should be remarked that the above taxonomy systematically classifies ADS technology. Companies can always deviate from this taxonomy and add additional, dedicated conditions when promoting and/or selling contracts for their ADS-equipped cars. For example, Tesla in selling its Tesla-S model, which is a Level 3 ADS, inserted a condition in the contract that

[60] SAE, 24.

required the “driver to constantly remain vigilant and ready to regain operational (manual) control at *any* time.”[61]

II.3 Dual-mode driving: ADS Level 3

Of the six levels of automation for driving listed in the previous subsection, only Level 3 and up are captured under the label ADS. Level 3 ADS is currently becoming available on the market gradually, such as the Tesla-S model. Therefore, experience is also becoming available to further analyze and improve this technology. As outlined in the Introduction, for this thesis I am interested in the ethical aspects of using ADS.

The ethical research can be divided into two major parts. One is focused on embedding ethical reasoning inside ADS so it becomes capable to exercise moral judgment in its decision-making autonomously. This topic is treated in the next chapter and in Chapter V. The second is focused on the role of the human in the design *and* operation of ADS. This is because, for example, at Level 3 ADS the human still may have to play a key role in handling critical scenarios such as in case of failures of (components of) the ADS system, the vehicle, and/or the infrastructure. An important ethical element in this human-machine interaction is the handling of human responsibility, as discussed in the Introduction. As experience is available with Level 3 ADS, I will focus on this Level ADS in the analysis of the ethics of responsibility for ADS. This is done from Chapter VI onwards. In the literature,[62] Level 3 ADS is also called dual-mode ADS

[61] Giulio Mecacci and Filippo Santoni de Sio, “Meaningful Human Control as Reason-Responsiveness: The Case of Dual-Mode Vehicles,” *Ethics and Information Technology* 22, no. 2 (June 1, 2020): 111, https://doi.org/10.1007/s10676-019-09519-w, 'italics mine'.

[62] Mecacci and Santoni de Sio, “Meaningful Human Control as Reason-Responsiveness,” 103.

as ADS may revert to the driver in case of failures. Such dual-mode ADS considers two agents: the ADS (or its developers) and the human driver.

Chapter III

Programming Ethics in AI for ADS: A top-down approach

III.1 Introduction

The first part of sub-question one is stated in Chapter I, “How could we integrate ethical decision making in handling life-threatening traffic dilemmas?” One ‘classical’ ethical approach to address this question is to apply various normative ethical theories, such as ethics of deontology or teleology, to hypothetical life-threatening traffic dilemmas. This approach is followed in this chapter. It is called a “top-down approach” of machine ethics,[63] as it attempts via the application of normative ethical theories to devise rules that can be programmed into automated machines. If successful, machines could be programmed to act as moral agents that make their ‘own’ moral decisions.

The hypothetical scenarios are thought experiments invented by philosophers. The most famous one is the so-called trolley problem introduced by Philippa Foot as a “runaway tram.”[64] A more up-to-date version of this trolley problem,[65] along with two other thought experiments based on the work of Jeffrey Gurney, is presented in Section III.2.

Two generally used ethical theories are then applied and discussed to address these thought experiments. The utilitarian or teleological approach is discussed in Section III.3 and the deontological approach in Section III.4. I conclude this chapter with a critical analysis of this

[63] Wendell Wallach, “Robot Minds and Human Ethics: The Need for a Comprehensive Model of Moral Decision Making,” *Ethics and Information Technology* 12, no. 3 (July 2010): 243, https://doi.org/10.1007/s10676-010-9232-8.

[64] Philippa Foot, “The Problem of Abortion and the Doctrine of the Double Effect,” *Oxford Review* 5 (1967): 5–15.

[65] The dilemma in the most classical cases of the trolley-problem is to either save five people from being possibly hit by a tram by redirecting this tram to a sidetrack, causing one person on that track to be killed, or by pushing somebody in front of the tram so it will stop before hitting the five.

top-down approach, which stipulates the need for an alternative approach. Such an alternative is the bottom-up approach developed in Chapter V. Here use will be made of some insights of the structure of responsibility of Bonhoeffer, which is reviewed in the next chapter.

III.2 Ethical Thought Experiments of Life-threatening Traffic Dilemmas

Philosophers have presented several thought experiments to analyze ethical dilemmas in traffic scenarios with ADS. See, for example, the Internet site "the Moral Machine" of the Massachusetts Institute of Technology (MIT).[66] This site presents a graphical traffic dilemma that a self-driving car could face. Visitors to the site can offer their opinion on which of two possible outcomes should be taken in each presented traffic dilemma. These dilemmas test nine separate factors that evaluate the visitors' preference of crashing into women versus men, young versus elderly, pedestrians versus jaywalkers, low-status versus high-status individuals, or saving few versus saving more lives. The dilemmas are all variants of the trolley problem presented in Subsection III.2.3. Over two years, millions of people have provided their opinions. From these results, the organizing MIT team derived some consistencies in preferences. These include the following: sparing humans over animals, saving more lives rather than fewer, and favoring children over adults.[67]

These preferences may help policymakers to create laws for self-driving cars and to determine a possible normative ethic on how to address these dilemmas in formal ethical theories

[66] Moral Machine, "Moral Machine," Moral Machine, accessed June 30, 2020, http://moralmachine.mit.edu.

[67] James Vincent, "Global Preferences for Who to Save in Self-Driving Car Crashes Revealed," *The Verge*, October 24, 2018, https://www.theverge.com/2018/10/24/18013392/self-driving-car-ethics-dilemma-mit-study-moral-machine-results.

should be used. This is done in Subsections III.3 and III.4 for the dilemmas presented in this section.

III.2.1 The Shopping Cart Problem

I take the variant of this problem, as stated by Jeffrey Gurney:

> After pulling into the parking lot of a grocery store, an autonomous vehicle's brakes stop working. Directly in front of the autonomous vehicle is a mother pushing a baby carriage. To its left is an overloaded shopping cart, and to its right is the grocery store. Assuming that any of the choices would be capable of stopping the autonomous vehicle, what should the autonomous vehicle do?[68]

Recall that we seek a moral approach and not a legal one to address this dilemma, as, from a legal perspective, the manufacturer of the brakes might very well be culpable. To address this problem from a moral point of view, crucial information might be missing, such as whether the baby carriage contains a baby. In moral analysis, these additional aspects are the *framing* of the problem.[69] When it is not clear whether there is a child in the baby carriage, one has to estimate what is inside. As this information may be lacking, a crucial aspect in moral decision-making is to deal with such uncertainty or lack of information. In prescriptive ethics, uncertainty is, in general, neglected; the classical way of applying ethical theories to moral dilemmas is by assuming that all ethically relevant conditions are known. This would then mean that in the above formulation of the shopping cart problem, one knows that a baby is in the baby carriage.

[68] Jeffrey Gurney, "Crashing into the Unknown: An Examination of Crash-Optimization Algorithms Through the Two Lanes of Ethics and Law," *Albany Law Review* 79, no. 183 (March 2016), 195. https://papers.ssrn.com/abstract=2622125.

[69] Gurney, 215.

III.2.2 The Motorcycle Problem

This comprises the following traffic dilemma:

> An autonomous vehicle encounters a situation in which it must strike one of two motorcyclists. To the vehicle's front-left is a motorcyclist who is wearing a helmet. To the vehicle's front-right is a motorcyclist who is not wearing a helmet. Which motorcyclist should the autonomous vehicle strike?[70]

This problem is perceived differently, depending on countries' laws. For example, only nineteen states of the US have motorcycle helmet laws that require all riders to wear helmets. However, in the EU, wearing an EU-approved helmet is mandatory for motorcyclists. According to a research report of the US National Highway Traffic Safety Administration, wearing a helmet "reduces the risk of fatality by 22 to 42%, and reduces the risk of brain injury by 41 to 69%."[71] These figures show that the motorcyclists not wearing a helmet have a higher chance of injury or death in a collision with the autonomous vehicle. This is again an example of additional information that may be crucial in a moral analysis of this dilemma.

III.2.3 The Trolley Problem

From Jeffrey Gurney, I cite the following variant:

> An operator is driving her autonomous car in manual mode and is in control of the vehicle. Whether intentionally or not — she could be homicidal or simply inattentive — she is about to run over and kill five pedestrians. Her car's crash-avoidance system detects the possible accident and activates, forcibly taking control of the car. To avoid this disaster, the car swerves into the only direction it can — say, to the right. However, on the car's right is a single pedestrian who the car strikes and kills.[72]

[70] Gurney, "Crashing into the Unknown: An Examination of Crash-Optimization Algorithms Through the Two Lanes of Ethics and Law," 197–98.

[71] Nat'l Highway Safety Admin., U.S. Dep't of Transp., *Countermeasures That Work: A Highway Safety Countermeasure Guide for State Highway Safety Offices*, Countermeasures that Work (Chapel Hill, NC: University of North Carolina, 2011), Chapter 5–1, http://www.nhtsa.gov/staticfiles/nti/pdf/811444.pdf.

[72] Gurney, "Crashing into the Unknown," 205–6.

In this case, the ethical question is on which moral grounds the programmer of the autonomous system decides to kill one person over five. On legal grounds, as the US law does not typically require people to act, the manufacturer would not be held responsible for the death of the five pedestrians when it did not take control to swerve into them instead of into the single pedestrian. However, "if the autonomous technology takes control of the vehicle to save five lives and ends up killing one person, the manufacturer is civilly and perhaps even criminally responsible for the death of the person killed."[73]

III.3 Teleological Ethics for Addressing the Ethical Thought Experiments

From teleological or consequential moral theory, utilitarianism is often considered when addressing moral dilemmas with autonomous machines because it allows one to translate every decision into a moral "calculation" of the gain or loss indicated by the utility measure of the consequences related to that decision (or action). Such a measure, for example, could be the cost of damage involved as the consequence of the decision taken.

Within utilitarianism, there is an important distinction between act utilitarianism and rule utilitarianism. Under act utilitarianism, an act is right if and only if it produces the greatest gain or the least cost for the individual taking the decision. For rule utilitarianism, "an act is right if and only if it is required by a rule that is itself a member of a set of rules whose acceptance would lead to the greater utility for society than any available alternative."[74] Rule utilitarianism

[73] Gurney, 207–8.

[74] Gurney, 211.

is typically used in a pluralistic democratic society. Here, usually, “moral debates … focus on the social consequences of the proposal in question,” before taking a legislative decision.[75]

III.3.1 Application of Utilitarianism to the Shopping Cart Problem

When considering the situation as defined in the Shopping Cart Problem of Subsection III.2.1, act utilitarianism would estimate the cost of the damage caused in each of the possible options. Consider the collision with the grocery store: this would yield damage to the store, possibly also bodily harm to shoppers and employees inside the building, the autonomous car, and its occupants. The cost may be expressed in monetary terms. Such an estimate may be difficult to make since it is impossible to see what is behind the wall of the grocery store. The cost of damage is greatly different when hitting a primary structure of the building where human beings were present compared to hitting a storage place without human beings and not containing any vital building structure.

For the second option, namely for the car to collide with the shopping cart, the cost of damage will include the destruction of the shopping cart and its content, as well as the damage to the autonomous vehicle and possibly to (some of) its occupants. Furthermore, estimates are needed, such as the cost of the damaged products in the cart, as well as the damage to the cart, the car, and its occupants.

The third option, colliding with the baby carriage, would cause major damage to the baby carriage and the possible grievous harm to the baby it contains. The damage to the autonomous vehicle and its occupants might be minimal in this case.

[75] Patrick Nullens and Ronald T. Michener, *The Matrix of Christian Ethics: Integrating Philosophy and Moral Theology in a Postmodern Context* (London: IVP Books, 2010), 52.

As the third option will probably cause the death of the baby, the cost of that option in terms of human life is the highest of the three cases. This would probably make the costs of the second option the lowest, which, according to act-utilitarianism, will be the best option. When this is the case, rule utilitarianism will arrive at the same conclusion, as society would accept as a rule that one should opt for causing minimal damage in case damage cannot be avoided. Rule utilitarianism would also rule out the possible killing of the baby.

III.3.2 Application of Utilitarianism to the Motorcycle Problem

Taking into consideration the figures about the reduced risk of death and/or brain injury as listed in Subsection III.2.2 would lead to the estimate that less damage would likely result from the motorcyclist wearing a helmet. Under the assumption that the damage of the impact on the autonomous vehicle and motorcycle are comparable, act utilitarianism, selecting the least damage-causing scenario, would opt for hitting the helmet-wearing motorcyclist.

On the other hand, under rule utilitarianism, the reduced risks would make a society opt for the rule that motorcyclists should wear a helmet. To stimulate this fact, a rule to promote maximal happiness could be to protect motorcyclists wearing a helmet. Therefore, rule utilitarianism, in the application of this rule, would elect to hit the motorcyclist who is not wearing a helmet.

The decision of the act utilitarianism would be seen as “unfair for someone solely because she was responsible.”[76] Therefore, this example clarifies that only considering act

[76] Gurney, “Crashing into the Unknown,” 198.

utilitarianism to ‘program an automated vehicle’ when handling accidents “may not take into account other important society values, such as fairness.”[77]

III.3.3 Application of Utilitarianism to the Trolley Problem

From a utilitarian viewpoint, both act and rule-based, the choice would be simple: the automated vehicle “would take control of the vehicle and kill the one person — five lives are better than one life.”[78]

The fact that autonomous driving technology is geared towards improving safety would be an argument for the automated vehicle to act in the case of the described trolley problem. However, whether to act in the way suggested above is a matter of moral deliberation.

III.4 Deontological Ethics for Addressing the Ethical Thought Experiments

The word deontology is derived from the Greek words δεον and λογος, meaning respectively “duty, obligation,” and “reason, study.”[79] This etymology captures the essence of deontology: it is, indeed, the study of the nature of duty and obligation. From the many types of deontological ethical theories, this chapter focuses on the classical Kantian definition. For the Enlightenment philosopher Immanuel Kant, the rightness of an action is not determined by its consequences but by the motives of the underlying action. Kant’s theory of ethics is based on “absolute rules — better known as ‘categorical imperatives’.”[80] A “categorical imperative imposes itself on every

[77] Gurney, 198.

[78] Gurney, 215.

[79] H. G. Liddell et al., *A Greek-English Lexicon*, 9th edition with Revised Supplement (Oxford, New York, NY: Oxford University Press, 1996).

[80] Gurney, “Crashing into the Unknown,” 218.

moral and responsible person, regardless of the person's self-interest or the possible consequences."[81] According to Kant, an absolute rule must comply with three formulations to qualify as being morally good.[82] The first formulation is "Act only on that maxim by which you can, at the same time, will that it should become a universal law."[83] An example of this formulation is lying. If it is generally accepted that everyone lies, no one will pay attention to what is said. This would be self-defeating and, therefore, "lying could not be willed into a universal law, and, thus, lying is always immoral."[84] The second formulation is "Act in such a way that you treat humanity, whether in your own person or in any other person, always at the same time as an end, but never merely as a means."[85] This formulation "only prohibits people from treating others *merely* as a means to their end."[86] The third formulation is "the idea of the will of every rational being as well that legislates moral law."[87] This formulation "represents Kant's belief that an act must be done out of a sense of duty — and not out of inclination — for the act to have moral worth."[88] The rational being formulates its own laws and then submits to itself.

In the following subsections, I apply Kant's absolute rule to the three hypothetical dilemmas formulated in Section III.2. In all these cases, the autonomous vehicle cannot obey the

[81] Nullens and Michener, *The Matrix of Christian Ethics*, 106.

[82] Nullens and Michener, 106.

[83] Immanuel Kant, *Groundwork for the Metaphysics of Morals*, ed. by Thomas E. Hill Jr and Arnulf Zweig, Oxford Philosophical Texts (Oxford, New York, NY: Oxford University Press, 2002), 222.

[84] Gurney, "Crashing into the Unknown," 218.

[85] Kant, *Groundwork for the Metaphysics of Morals*, 230.

[86] Gurney, "Crashing into the Unknown," 218.

[87] Kant, *Groundwork for the Metaphysics of Morals*, 232.

[88] Gurney, "Crashing into the Unknown," 219.

third formulation. However, the programmer of the crash handling software can act autonomously and he or she, therefore, can act as a rational agent. This makes it possible to program an ADS according to Kantian ethics.[89]

III.4.1 Application of Kant‘s Deontology to the Shopping Cart Problem

The striking of the baby carriage would be excluded from the options when the carriage contains a living baby. This is due to the following two reasons: first, the baby has inherent value, unlike the shopping cart and the grocery store, assuming that, in the last two cases, no human beings will be hurt, and, second, Kant’s second formulation prohibits treating the baby as merely a means to an end of the autonomous vehicle and its occupants.

III.4.2 Application of Kant’s Deontology to the Motorcycle Problem

In this dilemma, Kant’s deontology does not provide an action rule when making a choice of which motorcycle to hit, when, in both cases, the killing of the motorcyclist is a possibility. Moreover, when, in both cases, a lifelong injury would be the consequence of the collision, as cannot be excluded from the given statistics in Subsection III.2.2, Kant’s deontology cannot make a choice.

III.4.3 Application of Kant’s Deontology to the Trolley Problem

Similar to the case discussed in Subsection III.4.2, Kantian reasoning does not help to make a choice here. The choice to kill one person instead of five is excluded by the second formulation, as this would entail treating one person as merely a means to an end for the five persons.

[89] Gurney, 220.

III.5 Evaluation of the Top-down Approach

The philosophical top-down approach to devise decision-making algorithms for handling crashes with autonomous vehicles is inadequate, for several reasons. First, as we have seen, the different ethical theories provide different answers for a particular traffic (accident) dilemma. The discussion among philosophers has not been conclusive and it continues.[90] Second, most of the top-down approaches consider the handling of the accident from the perspective of a single autonomous vehicle only. The interaction and negotiation between different autonomous vehicles are only recently being addressed, and are discussed in Chapter V. A further analysis of the inadequacy to ethically resolve real accident dilemmas is based on the work of Nyholm and Smids.[91] In addition to the already mentioned shortcomings, they add a third one, namely that the top-down approach only considers a small number of morally relevant factors. This is very different from the real-world ethical reality of cars with ADS, where it is generally not possible to ignore the complexity and reality of the decision scenario without substantially changing the problem. An example of this occurs in the Shopping Cart Problem, where we can choose whether to assume that the baby carriage contains a baby. This shortcoming is also related to the fact that the considered traffic dilemmas are deterministic and consider only known precise scenarios. This is contrary to real-life accident scenarios where such knowledge is lacking and should be assumed in a probabilistic risk analysis framework.

[90] Jason Millar, "Ethics Setting for Autonomous Vehicles," in *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, ed. by Patrick Lin, Abney Keith, and Jenkins Ryan (Oxford, New York, NY: Oxford University Press, 2017), 23.

[91] Sven Nyholm and Jilles Smids, "The Ethics of Accident-Algorithms for Self-Driving Cars: An Applied Trolley Problem?," *Ethical Theory and Moral Practice* 19, no. 5 (1 November, 2016): 1275–89, https://doi.org/10.1007/s10677-016-9745-2.

The two major short shortcomings of not considering the abilities of the other ADS and/or neighboring traffic, as the responsible other, as well as the lack of realism, are addressed in Chapter V. This chapter concludes with an outline of a discussion on the value of the analysis of trolley-like dilemmas, such as those discussed in this chapter.

The paper of Keeling mentions two possible areas where the analysis of trolley-like dilemmas of crash scenarios is of practical relevance for developing ethics for ADS.[92] The first is that abstraction helps us to “show that theories of what matters morally in ADS collisions are false.”[93] The example Keeling gives is that it is false only to consider avoiding harm as this might require the ADS to kill somebody to avoid harm to many others. On the more positive side, abstraction can help us to instantiate those acts in practical dilemmas that are morally relevant. For example, I mention the “‘doctrine of double effect’ to distinguish between foreseeable killings and intentional killings.”[94] Under this doctrine, it is morally acceptable to incidentally kill one person to save five others, but the intentional killing of one person to save five others is not. The first situation holds in the given definition of the trolley problem, and it would favor the killing of the one instead of the five human beings. The second situation occurs when a medical doctor has to decide to take the life of one person to save five others. Thus, while the analysis of trolley-like crash dilemmas will not provide moral decision algorithms for real-time use in ADS, it does provide insight that will help philosophers and engineers work together by better understanding one another. Secondly, it highlights shortcomings that stimulate

[92] Geoff Keeling, “Why Trolley Problems Matter for the Ethics of Automated Vehicles,” *Science and Engineering Ethics* 26, no. 1 (1 February, 2020): 293–307, https://doi.org/10.1007/s11948-019-00096-1.

[93] Keeling, 297.

[94] Gurney, “Crashing into the Unknown,” 220.

further research that brings engineers and philosophers together to devise improved solutions. For example, one can consider the uncertainty about the facts in the traffic dilemma, such as the uncertainty that the autonomous vehicle will hit the baby carriage when the latter is pushed harder after sounding the horn of the autonomous vehicle. One can consider this uncertainty by using so-called *expected utility maximization*.[95] This mathematical framework allows us to define different worlds. For example, in the baby carriage case, there might be two 'worlds':[96] one where the autonomous car hits the carriage and one where it does not. In each world, an action can result in a particular outcome of the utility function. The expected utility is then the weighted sum of these utility outcomes in these different worlds, and the weights are taken to be the probability of being in that world. A utilitarian approach would aim to maximize this expected utility function. I refer to Keeling's work for more algorithmic details.[97]

To overcome the two major drawbacks of the top-down approach discussed in this chapter, namely to include the 'other' and to consider the realism of the situation in the ethical deliberation, I use insights of the ethics of responsibility of Bonhoeffer in Chapter V. First, the next chapter reviews Bonhoeffer's structure of the responsible life. In addition to this classical pioneer in ethics, I also review the work of Richard Niebuhr. The latter is used in Chapter VII to develop a new methodology to enhance the human-machine interaction from the perspective of reducing the responsibility gap for level 3 ADS.

[95] Keeling, "Why Trolley Problems Matter for the Ethics of Automated Vehicles," 300.

[96] Keeling, 300.

[97] Keeling, 298–300.

Chapter IV

Theological Ethics of Responsibility

IV.1 Introduction

As demonstrated, a major shortcoming of the top-down approaches to program ethics for ADS is the lack of realism of the considered driving scenarios. The hypothetical traffic (accident) scenarios only occur in extremely rare situations; furthermore, relevant information on the situation before the accidents, as well as the exactness of information needed, makes this approach, restricted to textbook examples, far removed from reality.

As "ethical reflection demands social action and difficult decision in real time and in real places,"[98] this chapter investigates two alternative approaches of ethical theories that consider the actuality of the ethical dilemma in real time. In both theories, the responsibility of the human being(s) involved in the ethical dilemma is crucial. In the context of this thesis, the human beings considered are those that design and operationally use automatic systems such as ADS because it is assumed that such autonomous systems do not yet have the same moral agency as human beings. Many others also make this assumption, such as Michael Horowitz, Adjunct Senior Fellow of the US Technology and National Security Program.[99]

The first alternative ethical theory is the theory of responsibility offered by the theologian Dietrich Bonhoeffer in his well-known primer entitled *Ethics*. The second alternative is the

[98] Nullens and Michener, *The Matrix of Christian Ethics*, 23.

[99] Michael C. Horowitz and Paul Scharre, *Meaningful Human Control in Weapon Systems: A Primer* (Washington D.C.: Center for a New American Security, 2015), 8, https://www.jstor.com/stable/resrep06179.

theory of responsibility offered by H. Richard Niebuhr. Before reviewing these two alternative ethical theories, a section is presented that motivates their selection. The work of Bonhoeffer and Niebuhr is then reviewed respectively in the subsequent two sections. These two sections are organized in the following manner. First, I start with some relevant comments on their biographies, which is followed by a summary of their contributions to the ethics of responsibility. Thereafter, two subsections briefly review the relevance of their contribution and review criticism of some of their contributions relevant to this thesis. In the final section of this chapter, the potential of the two theories is discussed concerning improving the treatment of ethics of responsibility for ADS.

IV.2 Motivation for using Bonhoeffer's and Niebuhr's work on responsibility

Bonhoeffer and Niebuhr were pioneers in developing alternatives from a theological perspective for the mainstream normative ethical frameworks of teleological and deontological ethics of their day. The sociologist Max Weber was the first to introduce the terminology 'ethics of responsibility' in his famous lecture "Politics as vocation."[100] While Bonhoeffer refers to Max Weber, he does not refer to that terminology[101] because for Bonhoeffer and Niebuhr, the concept of responsibility itself and the responsible self, respectively, attracts their interest, and not the concept of an ethic or ethics.

The works of Bonhoeffer and Niebuhr have inspired contemporary theologians to develop a Christian ethics of responsibility. Two important contributors to this endeavor are the

[100] Max Weber and Ronald Speirs, "Political Writings," (Cambridge: Cambridge University Press, 1994), 309–69, https://doi.org/10.1017/CBO9780511841095.

[101] Huber, "Ethics of Responsibility in a Theological Perspective," 193.

American Catholic theologian William Schweiker and the German Protestant theologian Wolfgang Huber. This section briefly elaborates on their contributions and, thus, highlights the selection of Bonhoeffer and Niebuhr as dialogue partners.

William Schweiker formulates his theological ethics of responsibility as follows: "in all actions and relations we are to respect and enhance the integrity of life before God."[102] However, as he remarks, his account "does not manage to incorporate possible validity of other beliefs into his ethics of responsibility."[103] This is contrary to Niebuhr's openness towards including other beliefs in his ethics of responsibility. The same can be said about Bonhoeffer's work on responsibility because Bonhoeffer's "desire to speak to the world — throughout his work he developed multiple concepts by which to overcome what he perceived to be a false distinction between 'Christian' and 'secular' concerns."[104]

Bonhoeffer's desire to speak to the world has certainly contributed to the vast interest of both Christian and non-Christian scholars who use his work to analyze a wide variety of moral issues. Examples include the political aftermath of apartheid in South Africa[105] and the moral debate on bioscience, environmental ethics, and the ethics of resistance, all mentioned in *Bonhoeffer's Christocentric Theology and Fundamental Debates in Environmental Ethics*.[106] A

[102] William Schweiker, *Responsibility, and Christian Ethics* (Cambridge: Cambridge University Press, 1999), 2.

[103] Schweiker, 24.

[104] Steven C. van den Heuvel, *Bonhoeffer's Christocentric Theology and Fundamental Debates in Environmental Ethics* (Eugene, ORE: Pickwick Publications, an Imprint of Wipf and Stock Publishers, 2017), 2.

[105] Esther D. Reed, "The Limits of Individual Responsibility: Dietrich Bonhoeffer's Reversal of Agent-Act-Consequence," *Journal of the Society of Christian Ethics* 37, no. 2 (Fall/Winter 2017): 39–58.

[106] Van den Heuvel, *Bonhoeffer's Christocentric Theology and Fundamental Debates in Environmental Ethics*, 2.

further illustration of that versatility is presented at end of Section IV.3.8 on the ethics of leadership.

Huber's attempt to develop a Christian ethics of responsibility is described in the article "Towards an Ethics of Responsibility"[107] This description is characterized by four structural elements.[108] While the elements "foundation in a relational anthropology" and "correspondence to reality" have much in common with Bonhoeffer's and Niebuhr's work on responsibility, Huber's third structural element understands ethics of responsibility as teleological ethics. This makes his work of less interest to this thesis compared to the work of Bonhoeffer and Niebuhr on ethics of responsibility.

Therefore, the openness of both Bonhoeffer's and Niebuhr's work on responsibility towards engaging with secular viewpoints, as well as their pioneering works on developing alternatives for the normative ethical frameworks of deontology and teleology, make them the preferred dialogue partners for engaging in a dialogue with the secular approaches that analyze the ethics of responsibility for ADS in this thesis.

IV.3 Bonhoeffer on the Structure of Responsible Life

IV.3.1 Introduction

The section "Structure of Responsible Life" in "History and Good [2]"[109] is one of the most widely referenced parts of Bonhoeffer's *Ethics*, his unfinished magnus opus. In this section, I outline the four vertices of the structure of responsibility described by Bonhoeffer. The first two

[107] Wolfgang Huber, "Toward an Ethics of Responsibility," *Journal of Religion* 73, 4 (October 1, 1993): 573–91, https://doi.org/10.1086/489259.

[108] Huber, 580–89.

[109] Dietrich Bonhoeffer, *Ethics*, 246.

affirm the bond of the responsible person to both God and neighbor. These two take the form of "vicarious representative action" (*Stellvertretung*) and "accordance with reality" (*Wirklichkeitsgemäßheit*). The second two affirm the freedom of the responsible person's life and take on the form subsequently of "willingness to actively embrace guilt" (*Bereitschaft zur Schuldübernahme*) and the "free venturing of a concrete decision" (*Wagnis der konkreten Entscheidung*).

The above four vertices are to be framed within Bonhoeffer's aversion to the dominating ethical thought that regards the individual as possessing an absolute criterion about what is good or evil and who can make a clear distinction between these two. According to Bonhoeffer, the notion of good is not obtained by abstraction from life, but precisely when humans immerge themselves in it. Life, in this instance, is not to be regarded as a thing or a concept but the person of Jesus Christ (John 14:6). "In our relation to Jesus Christ, the 'Yes' of creation, reconciliation, and redemption, and the 'No' of judgment and death over life that has fallen away from its origin, essence and goal" are being recognized as our life.[110] Bonhoeffer thus concludes that humanity can only be understood in Jesus Christ. In Him, our encounter with other human beings and God is subjected to the same yes and no. Responsibility is now our answer to the life of Jesus Christ. According to Bonhoeffer, this is contrary to the partial answers offered by ethical considerations based on usefulness or certain principles.

This section begins with a brief biographical sketch of Bonhoeffer to highlight several relevant issues in his life to better understand his contribution to this thesis. I then summarize Bonhoeffer's four vertices of the structure of responsibility based on "History and Good [2]."

[110] Bonhoeffer, 251.

The final two subsections present the relevance of the structure of responsibility in Bonhoeffer‘s oeuvre as well as its contemporary use, which is followed by a discussion on the criticism of the key components of that structure.

IV.3.2 Biographic Sketch of Dietrich Bonhoeffer

Dietrich Bonhoeffer (1906 - 1945) grew up in a Bourgeois family of intellectuals that observed the Christian rituals pro forma. His grandparents provided examples of activism to protest the social injustice of different German governments before the Hitler regime.[111] Much about Bonhoeffer’s life is revealed in the biography of Eberhart Bethge, who was Bonhoeffer’s closest friend and brother-in-law.

The Second World War had a major impact on Bonhoeffer. When Germany was in the fierce grip of the Nazi regime, he wrote a book on ethics partly motivated by the German Church’s failure to publicly criticize the tyranny of that regime. For Bonhoeffer, the reluctance of the confessing German church made the issue of ethical responsibility perhaps the most demanding concern.[112]

Bonhoeffer’s development of ethics was influenced by two facts. The first was the recognition of the impotence of traditional virtue and principle ethics. These normative ethical theories had contributed, according to Bonhoeffer, to the failure of the Christian German church to criticize the German Nazi regime for its atrocities and caused many German officers not to

[111] Peter Frick, “Dietrich Bonhoeffer: Engaging Intellect – Legendary Life,” *Religion Compass* 6, no. 6 (June 28, 2012): 310, https://doi.org/10.1111/j.1749-8171.2012.00357.x.

[112] Frick, “Dietrich Bonhoeffer,” 314.

take part in active resistance against that regime because of their oath of loyalty to Hitler.[113] The second was Bonhoeffer's involvement in the conspiracy against Hitler for which the Gestapo arrested him on 5 April 1943. "Up to that moment, Bonhoeffer had been working on the manuscript of *Ethics*."[114] Bonhoeffer's involvement in tyrannicide combined with his support for pacifism posed a serious ethical predicament.

In prison, Bonhoeffer continued writing and penned his famous "Letters and Papers from prison."[115] Bonhoeffer was hanged on 9 April 1945, a month before the end of World War II.

IV.3.3 Vicarious Representative Action (*Stellvertretung*)

Vicarious representative action is the most evident vertex of the structure of responsibility. This action requires full devotion of one's own life to another person, and when a father is acting "on behalf of his children by working, providing, intervening, struggling, and suffering for them," he truly stands in their place.[116] Even when a father acts in a bad manner, this does not exempt him from his responsibilities to his children. All human life is a vicarious representation, even when we are alone because life exists only in Jesus Christ and even though He lived without the particular responsibility of a marriage or family, He accepted His responsibility to care for the poor and for humanity by dying on the cross. "Jesus was not the individual who pursued some personal perfection, but only lived to take on and bear the selves of all human beings. To live a

[113] Jens Zimmermann, "Virtue Ethics and Realistic Responsibility in an Age of Globalization," in *Handbook of Virtue Ethics in Business and Management*, ed. by Alejo José G. Sison, Gregory R. Beabout, and Ignacio Ferrero, International Handbooks in Business Ethics (Dordrecht: Springer Netherlands, 2017), 7, https://doi.org/10.1007/978-94-007-6510-8_54.

[114] Frick, "Dietrich Bonhoeffer," 314.

[115] Dietrich Bonhoeffer, *Letters and Papers from Prison*, (Minneapolis, MN: Fortress Press, 2010).

[116] Bonhoeffer, *Ethics*, 257–58.

responsible life like Jesus means to be selfless, and this creates unity within the responsible individual between the divine Yes and No.

Leading a vicariously responsible life presents two dangers. The first danger is absolutizing the self. That case ignores that only the selfless person can act responsibly. However, the result will be a violation of the right of others and tyranny. The second danger is absolutizing the other for whom responsibility is taken. This results in ignoring all other responsibilities and creates arbitrariness in your action. Both dangers deny the origin, essence, and goal of responsibility and make that responsibility a self-made, abstract idol.

Bonhoeffer goes on to argue that "[t]here is also a responsibility for things, conditions, and values, but only by strictly keeping in mind that the origin, essence, and goal of all things, conditions, and values is determined by Christ (John 1:4)."[117] Outside these limits, there is the danger that things dominate over people. In this respect, Bonhoeffer highlights that the values of truth, the good, the right, and the beautiful can sometimes receive so much attention that they become false idols. Then what might have been a responsible action may lead to a habitual obsession that hampers or even endangers being human. The second danger is the questioning of the usefulness of things, conditions, and values, as that will desecrate them. These dangers disappear, providing the "world of things ... its full freedom and depth only when it is seen oriented toward the world of persons in its origin, essence, and goal."[118] In other words, through Jesus Christ, things and values regain their orientation towards human beings as originally intended in the creation.

[117] Bonhoeffer, 259.

[118] Bonhoeffer, 260.

IV.3.4 In Accordance with Reality (*Wirklichkeitsgemäßheit*)

Action in accordance with reality means taking responsibility that is directed to concrete neighbors in their concrete reality. Such action occurs under Jesus Christ because everything that exists "receives its ultimate foundation and its ultimate negation, its justification and its ultimate contradiction, its ultimate Yes and its ultimate No" from Jesus Christ.[119] By this reconciliation of the world with Christ, the world is allowed to remain the world and human actions should occur in such a reconciled world, without the need to suffocate these actions under the burden of principles. When actions spring from principles, different forms of secularism or the teaching about 'autonomous spheres of life' ensue; or, on the other hand, it leads to religious enthusiasm. In both cases, the result is the destruction of the world reconciled in Christ with God.

"[H]uman beings are placed in a position of concrete and [are] thus limited, i.e. created, a responsibility that recognizes the world as loved, judged and reconciled by God, and acts accordingly."[120] Therefore, human actions, subjected to reality, are limited by the fact that we are creatures. The consequences of this are manifold. Bonhoeffer derives the following helpful guidelines to determine your action according to reality,[121] as clustered in the following three groups:

1. Taking your own finiteness into account, by

— being aware that you cannot create the conditions of your action yourself but find yourself already placed within them.

[119] Bonhoeffer, 261–62.

[120] Bonhoeffer, 267.

[121] All elements of the this list are taken from Bonhoeffer, 267–68.

— acknowledging certain limitations from both the past and the future that you cannot leap over.

— knowing that your responsibility is not infinite but limited, and within these limits lies the whole reality.

2. Being willing to make trade-offs, by

— being not only concerned with good intension but also with good outcomes.

— seeking to understand the entire given reality in its origin, essence, and goal, and seeing it under the divine Yes and No.

— stipulating your objective, not as an application of some kind of limitless general principle but requiring in the given situation to observe, weigh, evaluate, and decide, and to do all that with limited human understanding.

— having the courage to look into the immediate future.

— seriously considering the consequences of our actions.

3. Conducting introspection and extrospection, by

— seriously examining your own motives and your own heart.

— not attempting to revolutionize the world but, at the given place, to consider reality and do what is necessary.

— asking what is possible since you cannot take the final step right away.

— not being blind.

— regarding other people you encounter as responsible. This may require making them aware of their responsibility and that their responsibility may limit yours.

The difference between humans acting responsibly in accordance with reality and action based on preconceived principles entails, in the first case, that one has to completely surrender judgment of that action to God, while in the second, justification is already based on the self-defined principle. Therefore, the latter renounces any knowledge about the ultimate justification. Instead, living in line with reality requires living in full dependence on grace, ignoring one's own goodness and evil. Responsible action in this way places this action in "the hands of God and live by God's grace and judgment."[122]

In our responsibility to the world of things, it is first important to discover the "intrinsic law" of those things.[123] Such discovery will become more difficult the closer the object is to our human existence.[124] For example, technical laws are easier to discern than laws characterizing human relations. Second, when taking the appropriate action, one must consider how the subject matter is related to the person. Such relationships might be distorted when the action is determined by either considering the status between the person and object to be independent, opposite, or standing side by side.

A final remark on the acting in accordance with reality concerns dealing with borderline cases, in other words, when we are confronted with an extraordinary situation of ultimate realities, which in politics is exemplified by war or breaking a treaty for the sake of one's own life necessities. In such borderline cases, responsibility is not constrained by any law, and a direct appeal should be made to the freedom that acting responsibly brings. Actions under the denominator of free responsibility call for accountability. This is the topic of the next section.

[122] Bonhoeffer, 269.

[123] Bonhoeffer, 271.

[124] This proximity between object and human is closely related to Emil Brunner's 'law of closeness', as outlined in his book *Truth as Encounter*. H. E. Brunner, *Truth as Encounter* (Philadelphia, PA: Westminster John Knox Press, 2000), 54-56.

IV.3.5 Willingness to Actively Embrace Guilt (*Bereitschaft zur Schuldübernahme*)

When Bonhoeffer wrote the "History and Good [2]", he was convinced that "everyone who acts responsibly becomes guilty . . . because Jesus Christ took the guilt of all humans on himself."[125]

Bonhoeffer sees two obstacles in human beings in becoming guilty for actions taken for the sake of others. First, there is the human conscience, which Bonhoeffer interprets very much in line with Kant's view as egocentric, as discussed in Section IV.4.7. This egocentricity follows from the authority that Bonhoeffer assigns to the human conscience in its possibility to refuse to sacrifice its integrity to any other good – the refusal to refrain from becoming guilty for the sake of another human being. Nevertheless, Bonhoeffer, suggests that the obstacle for taking on guilt lies in "one's own ego in its demand to be 'like God' — *sicut deus* — in knowing good and evil."[126] Human beings who, in faith, view the origin and goal of one's human conscience not in a law but in the living God and in the living human beings as they encounter them in Jesus Christ, believe that Jesus sets their conscience free for the service of God and their neighbor. Moreover, in this act, they will not shy away from entering into guilt for the other person's sake. This freedom in service to God aligns Bonhoeffer's view about the human conscience with that of Niebuhr, which is discussed in Section IV.4.7.

Second, this is in contrast to human conscience that seeks self-righteousness justified by following a principle. In this instance, Bonhoeffer attacks Kant's principle of truth-telling. According to this principle, you would have to answer 'yes' to the question of "whether you

[125] Bonhoeffer, *Ethics*, 275.

[126] Bonhoeffer, 277.

[would] offer refuge to a friend in your house who tries to escape the pursuit of a murderer."[127] For Bonhoeffer, such self-seeking righteousness of the human (through its conscience) is an obstacle to responsible action.[128] While Bonhoeffer recognizes that each person has his/her limit in "[t]he measure of guilt incurred in connection with a particular responsible action," a conscience that is freed in Jesus Christ, who now becomes the ultimate rather than the law, must, therefore, freely decide in favor of Jesus Christ.[129] In dealing with this incurred guilt, Bonhoeffer claims that those who freely accept responsibility are justified before the others because of the necessity of their action, while they are set free by their own conscience. In relationship to God, it is a matter of hope and grace.

This brings me to the last vertex of the structure of responsibility, namely, the freedom to make concrete decisions.

IV.3.6 Free Venturing of a Concrete Decision (*Wagnis der konkreten Entscheidung*)

People who act responsibly do that in their own freedom. They consider all existing circumstances in the activities related to people, conditions, and principles and, in so doing, acting in freedom means that they do not hide behind excuses, such as the lack of favorable conditions or trying to give meaning to the intended actions by first turning them into a principle. Someone acting in free responsibility can do well, not by knowing good and evil, but by

[127] Bonhoeffer, 279.

[128] This view of Bonhoeffer was revised later on in his writings, such as in the part of his *Letters and Papers from Prison* with the title "What Does It Mean to Tell the Truth?" Bonhoeffer, *Letters and Papers from Prison*. Section IV.3.9 elaborates further on this revision.

[129] Bonhoeffer, *Ethics*, 281–82.

surrendering his/her deeds to God, to let good take place without knowing it. This is because it is "God who looks upon the heart, weighs the deeds, and guides history."[130]

With all the burdens and pressures of daily life, one might have the impression that one is crushed by rules and regulations. To see the trees through the forest, it is then necessary to understand the relationship between free responsibility and obedience. In responsibility, obedience and freedom are linked. "Obedience without freedom is slavery."[131] By focusing on obedience only, one obtains Kant's ethic of duty. Then the obedient person is required to acquire his/her justification from within. The same occurs when focusing on freedom only, as "freedom without obedience results in arbitrariness."[132]

Responsible persons are bound by the first two vertices of the structure of responsibility to God and to his neighbor. Braveness is, nevertheless, required in finding justice, not in their human bonds or in their own freedom, but only in the One who placed the responsible person in the given challenging situation and who requires that person to act.

The next section describes the place where the structure of responsibility is exercised.

IV.3.7 The Place of Responsibility

Responsible life happens when one hears and responds to the call of Christ. This call is one's vocation and, as Jesus has reconciled the world, vocation is not to withdraw from the world. "Vocation comprises work with things and issues [sachliche Arbeit] as well as personal relations; it requires 'a definite field of activity,' though never as a value in itself but only in responsibility

[130] Bonhoeffer, 284.

[131] Bonhoeffer, 287.

[132] Bonhoeffer, 287.

to Jesus Christ."[133] The extent of one's vocation and its expansion in which one exerts or limits one's responsibility is only determined by the person's free response to the call of Christ. When determining the extent of one's vocation and the limits of one's responsibility, one can use self-evaluation criteria, even though these criteria cannot provide complete certainty. Examples of such criteria include the following: (a) to not let the limitation or expansion be based on principles, (b) be aware of biblical abuses, including possible exploitations, as in Luke 16:10; 19:17 to narrow one's scope of duty, or in 1 Pet 4:15 – to be reluctant to interfere with another person.

When expanding or limiting one's responsibility and vocation, one is confronted with the recurring problem to abide by the law of God as revealed in the Decalogue and in the divine mandates of marriage, work, and government, on one hand, and freedom, on the other hand. This now "threatens to introduce a contradiction into the will of God."[134] The free responsible person who breaks God's law should do that while considering the boundaries established by God in his laws with ultimate seriousness and when transgressing these laws in freedom; such a person does that to affirm what has been transgressed. Here, Bonhoeffer refers to examples of killing, lying, and seizing property during war for the sole purpose to restore the validity of life, truth, and property, respectively.

IV.3.8 Relevance of Bonhoeffer's Structure of Responsibility

The work on the structure of responsibility as summarized in the previous five subsections played a central role in Bonhoeffer's work. I highlight this for the vertex's vicarious

[133] Bonhoeffer, 292–93.

[134] Bonhoeffer, 296–97.

representation, as this vertex is “often regarded as one of the most decisive and innovative aspects of Bonhoeffer’s theology.”[135] The singular nature of the vertex on Guilt as well as Bonhoeffer’s correction based on the work of Matthew Puffer is discussed in the next subsection.

While the vertex’s vicarious representation is more developed in Bonhoeffer’s *Ethics,* it already appeared in earlier writings, such as in his doctoral dissertation *Sanctorum Communio* (SC),[136] which he defended at the age of 21. In this instance, Bonhoeffer situates responsibility as a fundamental social concept and not an individualistic one. While writing this dissertation, Bonhoeffer was influenced by personalist writers, such as Friederich Gogarten, Emil Brunner, and, most notably, Eberhard Grisenbach.[137] From the movement of personalism, Bonhoeffer drew the I-You relationship as a fundamental part of what it means to be human. Following this influence, Bonhoeffer recognizes that an individual becomes a person, in other words, an “I”, only in recognizing the other as a “You”. In SC, this I-You relationship is explicitly put forward along with the responsibility that such relationship brings forward in Christ and, as such, turns responsibility into a theological concept rather than a purely ethical one.

The chosen social philosophical and sociological perspective of Bonhoeffer’s SC was an attempt to understand the structures of the given reality of a church of Christ, as revealed in Christ. These insights can be correlated with (critical) traffic scenarios to emphasize collective thinking. In this translation, the aspect of the church community caring about the other, the

[135] Heuvel, *Bonhoeffer’s Christocentric Theology and Fundamental Debates in Environmental Ethics*, 218, based on the work of Abromeit, *Geheimnis Christi*, 268.

[136] Dietrich Bonhoeffer, *Sanctorum Communio: A Theological Study of the Sociology of the Church*, ed. by Clifford J. Green, trans. by Joachim Von Soosten, Reinhard Kraus, and Nancy Lukens, annotated edition (Minneapolis, MN: Fortress Press, 2009).

[137] Heuvel, *Bonhoeffer’s Christocentric Theology,* 216.

vulnerable one, and the goal of caring for the other in (critical) traffic scenarios will play a crucial role.

Bonhoeffer's SC is particularly useful because of the theological anthropology of collective-personhood. From SC, I deduce three features of this anthropology: (a) the Christological starting point, (b) Bonhoeffer's appeal to personalism, and (c) Bonhoeffer's concept of the collective person.[138] I briefly review these three features here, and they will be addressed later in Section VIII.5 in the ADS context.

The Christological starting point stems from Bonhoeffer's suspicion of all human efforts to describe the church — or humans — from a purely human (and, therefore, sinful) perspective. This is because, for Bonhoeffer, humans are incapable of escaping sin, which necessitates the need for divine intervention by Jesus Christ. This invites humans to align their creation and activities to the work of Christ – an alignment that calls for humility in what we do and create. While sin and failure will always interfere in human actions and creation, taking Christ as the starting point is reassuring as redemption can always be found.

Bonhoeffer's appeal to personalism was a reaction against "forms of German idealism which were viewed to be impersonal and therefore dehumanizing ... concepts of persons as mere expressions of large forces (contra Hegel)."[139] To counteract such anthropocentric views, Bonhoeffer adopts many elements of personalism, such as the I-Thou language that is mentioned in this subsection, which was also used by Niebuhr, as discussed in Section IV.4.3. In such relational encounters, Bonhoeffer appeals to the personal will to express the individual nature of

[138] Jeremy M. Rios, "Bonhoeffer and Bowen Theory: A Theological Anthropology of the Collective-Person and Its Implications for Spiritual Formation," *Journal of Spiritual Formation and Soul Care* 13, no. 2 (November 1, 2020): 185, https://doi.org/10.1177/1939790920915700.

[139] Rios, 186.

the human creature that is encountered and to make the human aware of being held culpable before God for sin. Anything more or less then considering the human will as a cause by which we sin, as well as by which we are saved, is considered by Bonhoeffer as fundamentally dehumanizing.[140]

The third element of Bonhoeffer's theological anthropology is the collective-personhood, which was unbroken before the Fall in its right relationship to God.[141] While the Fall has shattered our concept of the human collective-person, in Christ this collective-personhood is restored. It is important that Rios remarks that "[t]his collective-personhood does not entail loss of personal will, or collapse into a mystic mass," but it calls for a collective humanity such that "my live is lived with your life in community, and my life is further lived for your life in community (i.e. by means of goods, time, service, etc.).[142]

The vertex of vicarious representation also played a role in Bonhoeffer's writings on *Discipleship*.[143] However, in *Ethics*, "Bonhoeffer has shifted the emphasis from the qualitative difference between Christ‘s deputyship and man's doing good for his neighbor to the daily Christian and non-Christian life as a life that intimately partakes of Christ's deputyship at every turn and in a multitude of even unavoidable ways.[144] A final comment on the importance of the vertex of vicarious representation is that it also mentioned in his *Letters and Papers from*

[140] Rios, 186.

[141] Bonhoeffer, *Sanctorum Communio*, 62.

[142] Rios, "Bonhoeffer and Bowen Theory," 187.

[143] Dietrich Bonhoeffer, *Discipleship* (Minneapolis, MN: Fortress Press, 2015).

[144] Heuvel, *Bonhoeffer's Christocentric Theology*, 213, based on the work of Rasmussen, *Dietrich Bonhoeffer*, 40.

Prison,[145] though, in this case, I see a subtle but important shift in terminology as the word 'stellvertretung' is replaced by 'being-there-for-others'.[146]

In addition to highlighting the relevance of the structure of responsibility by recognizing its centrality in Bonhoeffer's oeuvre, I now briefly discuss how Bonhoeffer used, but also distinguished, himself from renowned philosophers of his time. Again, I limit myself to the vertex of vicarious representation only.

The link with the movement of personalism was already discussed above. In addition, Bonhoeffer built upon Max Weber's work on the ethics of responsibility. In *Ethics*, Bonhoeffer initially credits Weber, without reference to his introduced term 'ethics of responsibility' as remarked earlier in Section IV.2, but also Otto von Bismarck for their use of the word in a highly developed, ethical sense. However, he then sharply contrasts both of their views that the concept of responsibility should be interpreted in the biblical sense as "bearing witness before others (or giving account) of Christ."[147] This distinct view of the concept of responsibility provides a fundamental Christological focus to Bonhoeffer's work on responsibility.

A third factor highlighting the relevance of Bonhoeffer's structure of responsibility is that it is still relevant to contemporary writers. While Bonhoeffer's work is still influential today, as sociologists, philosophers, and Christians from different denominations still use it, for the purpose of this thesis, I focus on two contributions. This illustrates how Bonhoeffer's contributions were 'translated' in a contemporary context.

[145] Bonhoeffer, *Letters and Papers from Prison*.

[146] Heuvel, *Bonhoeffer's Christocentric Theology and Fundamental Debates in Environmental Ethics*, 213.

[147] Heuvel, 218.

Section I.2 mentions that the ethics of responsibility has both an anthropological and a theological basis. Bonhoeffer's structure of responsible life is defined by the togetherness of commitment and freedom. For Bonhoeffer, commitment emphasizes deputyship for those who are in need as well as realism in dealing with these challenges, while freedom accentuates the venture of accountability and the preparedness to become guilty.[148] Bonhoeffer's view of freedom not as a life in possibilities but as an encounter with reality is succinctly described in the following poem that he wrote in prison,

> Not always doing and daring what's random, but seeking the right thing.
> Hover not over the possible, but boldly reach for the real. Not in escaping to
> thought, in action alone is found freedom.[149]

From a contemporary anthropological perspective, the ethics of responsibility is seen as relational anthropology that understands persons as communicative beings, people who listen to their call and answer it. Sociologists, such as Ralf Dahrendorf, gave attention to the theological basis of that ethics by viewing the dialectic of commitment and freedom as "ligatures and options."[150] This small example demonstrates that the contemporary translation of Bonhoeffer's concepts removes the Christological dimension of his work.

Another illustration of the relevance of Bonhoeffer's structure of responsibility is its application to responsibility in leadership. This topic resembles the responsibility for ADS users as other agents may do most of the work while, in my view, as often stipulated in legal

[148] Huber, "Ethics of Responsibility in a Theological Perspective," 196.

[149] This is the English translation of Bonhoeffer's original poem that reads (in German): "Nicht das Beliebige, sondern das Rechte tun und wagen, nicht im Möglichen schweben, das Wirkliche tapfer ergreifen,
nicht in der Flucht der Gedanken, allein in der Tat ist die Freiheit" Dietrich Bonhoeffer, *Widerstand und Ergebung*, ed. by Christian Gremmels, Eberhard Bethge, and Renate Berthge, Dietrich Bonhoeffer Werke, vol. 8 (Gütersloh: Gütersloher Verlaghaus, 1998), 571.

[150] Huber, "Ethics of Responsibility in a Theological Perspective," 196.

agreements between ADS car manufacturers and users of ADS, these users remain responsible. For an interpretation of the ethical dimension, I make reference to a recent translation of Bonhoeffer's structure of responsibility by Van den Heuvel.[151] This paper does not view Bonhoeffer's structure as "a set of timeless principles that continually have to be re-embodied and reconfigured in new circumstances," but every new context requires a hermeneutical translation to distil from Bonhoeffer's four vertices the possible contributions to the ethical challenge at hand. In Van den Heuvel's paper, it is the field of contemporary leadership ethics.[152] In relation to the first vertex, an interesting fact highlighted in that paper was based on an ongoing debate: the way leadership looks at 'the other'. Such debates range from empowering capabilities of the leader giving rise to a sense of heroism to making 'the other' grow so dominantly that the self suffocates under boundless responsibility.[153] The dangers of viewing the other were already highlighted in the review of the vertex of vicarious representation in Section IV.3.3. Following Bonhoeffer, van den Heuvel asserts against such heroism that "human beings means beings in relation of responsibility — that is, that there is no autonomous Self as such."[154] In contrast to the other of the debate, based on Bonhoeffer's work van den. Heuvel further asserts that "the other should not be exalted so far above the self that the possibility of

[151] Steven C. van den Heuvel, "Leadership and the Ethics of Responsibility: An Engagement with Dietrich Bonhoeffer," in *The Challenges of Moral Leadership*, ed. by Patrick Nullens and Steven C. van den Heuvel, Christian Perspectives on Leadership and Social Ethics 2 (Leuven: Peeters, 2016), 111–25.

[152] Heuvel, 120.

[153] Heuvel, 121.

[154] Heuvel, 121.

negotiating their shared interests is closed off."[155] This danger was also highlighted in Section IV.3.3.

The second vertex is particularly relevant when addressing the speed by which new developments challenge leadership. These developments, such as globalization and increasing connectivity, open new business opportunities. On the other hand, they challenge ethics to deal with these new circumstances. Bonhoeffer's work can be used to advise a leader on how to deal with such seemingly 'unlimited' circumstances. Bonhoeffer's second vertex can contribute to this ethical challenge by providing limits in our encounter with reality, "namely the fallibility and the limitations (embodied both by God and other people) that come with being human."[156] To coping in a responsible manner as a leader within new circumstances, such as new technological developments, the articulation of limits is necessary for developing new modes of ethical thinking.

Applying Bonhoeffer's third vertex will challenge leadership not to try to remain ethically pure, but to act in a calculated manner, seeking to minimize guilt rather than eschewing it. Taking the complexity of decisions into consideration, the challenge here is to be able to conduct reasoned calculations that consider all available factors and getting these factors 'on the table' and correctly interpreted. This is vital when being confronted with uncertainties, as is the case in ADS decision-making based on uncertain data.

The fourth vertex underlines that, in the case of contemporary leadership, "a much greater emphasis needs to be placed on empowering both leaders and their subordinates for

[155] Heuvel, 121.

[156] Heuvel, 122.

responsible, ethical decision-making."[157] Such empowerment encourages those working under the leader(s) to make their own ethical choices, and it will increase their self-esteem and ethical performance.[158]

IV.3.9 Criticism and New Insights Concerning Bonhoeffer's Concept of Responsibility

The ongoing interest in Bonhoeffer's work has also resulted both in criticism as well as in new insights. Considering the scope of this thesis, I will restrict myself to addressing the criticisms leveled against Bonhoeffer's concept of vicarious representation; I will focus on new insights about his concept of 'willingness to embrace guilt'.

One critique on the first vertex is by Oswald Bayer. This criticism concerns Bonhoeffer's relation to Hegel's philosophy of religion. Bonhoeffer appreciates Hegel's insight that the self's liberation can only occur in relation to others, although Bonhoeffer rejects that relation with others can happen by self-reflection, to see the selves through the eyes of the other. Therefore, humans carry within themselves the potentiality for relationality, according to Hegel.[159] For Bonhoeffer, the I-You relationship is essential. Bayer criticizes Bonhoeffer's transforming of Christ's vicarious, representative action into an *abstractum*.[160] This would take away the concreteness of Bonhoeffer's essential Christological view. In line with this criticism, Bayer also charges Bonhoeffer that his Christology is based on his anthropology — and not the other way around.[161] Eberhart Jüngel supports this criticism and asserts that Bonhoeffer's Christ's vicarious

[157] Heuvel, 124.

[158] Heuvel, 124.

[159] Reed, "The Limits of Individual Responsibility," 45.

[160] Heuvel, *Bonhoeffer's Christocentric Theology and Fundamental Debates in Environmental Ethics*, 219.

[161] Heuvel, 219.

representation is an anthropological identifiable structure to make the mystery of Christ traceable, while, according to Jüngel, it should be the other way around: "The mystery of human existence is revealed in the encounter with Jesus Christ."[162] According to van den Heuvel, both criticisms result from a misunderstanding. Ann Nickson refutes the criticism of Jüngel and points out that "Bonhoeffer's theology and ethics are always grounded christologically; it is Christ the incarnate Son of God who defines true humanity, never vice versa."[163] Furthermore, the opposite accusation has been made, which states that Bonhoeffer's work is limited by Christian theology, or more specifically, by biblical revelation.[164] These accusations are a consequence of distinguishing Christ's vicarious representation from the social-ethical dimension of the vicarious representative vertex. However, "the social-ethical dimension of vicarious representation should be seen as a natural consequence of Christ's vicarious representation."[165]

A final new insight about Bonhoeffer's structure of responsibility, which is relevant to this thesis, is the work of Matthew Puffer on the third vertex of embracing guilt in borderline cases. In Section IV.3.5, I highlighted that the freedom of responsibility can introduce "a contradiction between the law of God and God's will."[166] Therefore, in borderline cases, where God's will and law are in conflict, "God wills that human persons recognize and break the laws God sets for them and, thereby, incur guilt."[167]

[162] Heuvel, 220, footnote 52.

[163] Heuvel, 220, footnote 53.

[164] Heuvel, 220.

[165] Heuvel, 221.

[166] Matthew Puffer, "Three Rival Versions of Moral Reasoning: Interpreting Bonhoeffer's Ethics of Lying, Guilt, and Responsibility," *Harvard Theological Review* 112, no. 2 (April 2019): 170, https://doi.org/10.1017/S001781601900004X.

[167] Puffer, 170.

Various scholars, such as Green, Rasmussen, and Pfeifer, have struggled with these provocative passages, as they appear to "provide warrants for lying, tyrannicide, and [the] attempted coup of Hitler."[168]

Contrary to the above-mentioned scholars, Puffer makes use of "extensive editorial work for the German critical edition of *Ethics* that shows that the chapter in that book 'God's Love and the Disintegration of the World' were not published in 1939 or 1940, but immediately after "History and Good [2]"."[169] As "History and Good [2]" breaks off where "God's Love and the Disintegration of the World" begins and "the close proximity of the distinct ways of relating God's law to God's will" in both these chapters of *Ethics,* Puffer concluded that "Bonhoeffer himself recognized as problematic the account of *Schuldübernahme* in his earlier discussion of lying and guilt quite quickly — insofar as it introduces a conflict between God's will and God's law — and that he, therefore, subsequently stopped advocating this position."[170]

After examining the complete corpus of Bonhoeffer, Puffer observes that the core message of the section "Active Embrace of Guilt" (*Schuldübernahme*) does not appear anywhere else in Bonhoeffer's entire corpus, despite its publication in "History and Good [2]". While this last viewpoint has received substantial attention in the ethics of Bonhoeffer, Puffer arrives – based on the analysis of how Bonhoeffer alternatively treats the issue of embracing guilt in his later publications – at the conclusion that "the conditions that gave rise to an active embrace of

[168] Puffer, 161.

[169] Puffer, 178.

[170] Puffer, 178, italics original.

guilt are no longer in place. God's will for human persons instead become synonymous with the doing of God's law."[171]

Therefore, one can ask how this later viewpoint of Bonhoeffer views the issue of guilt. Here, Puffer observes that in his final stage of life, Bonhoeffer was inspired by "people in the Old Testament [who] vigorously and often lie ... to the glory of God," and made him "recast truth-telling, and therefore lying, in a phenomenological register that relocates guilt and obviates the need to resolve a conflict of God's law with God's will."[172]

For example, in the case study in "which a child gives false information in response to his teacher's query regarding his father's drunkenness," treated in the essay "What does it mean to tell the truth?", "a novel ethic of lying unfolds that clearly contradicts the presentation in 'History of Good [2]'."[173] Now for Bonhoeffer, "the child's speech corresponds to the truth that God has mandated the family as a sphere into which the teacher may not intrude," concluding that "[i]t is exclusively the teacher who is guilty of lying."[174] Bonhoeffer wrote about this example in the above essay during the months that he was interrogated and tortured for his alleged involvement in the conspiracy against Hitler.[175] The child in Bonhoeffer's mentioned essay must be seen as a defense against his lies against his interrogators as they did not present a legitimate authority, as the teacher in the child's example, and, therefore, these interrogators were not entitled to question him. As a consequence, he was not obliged to answer truthfully.

[171] Puffer, 176.

[172] Puffer, 182.

[173] Puffer, 174–75.

[174] Puffer, 175.

[175] Scott Paeth, "The Responsibility to Lie and the Obligation to Report," *Journal of Business Ethics* 112, no. 4 (January 2013): 563.

IV.4 Niebuhr’s Responsible Self

IV.4.1 Introduction

Helmut Richard Niebuhr (1894 - 1962) was raised in a German, Evangelical, immigrant family that lived in Missouri, US. He wrote his dissertation entitled *Ernst Troeltsch’s Philosophy of Religion,* in 1924. Niebuhr was an influential scholar at Yale University, where he taught from 1931 until his death. At Yale he specialized in theology and Christian ethics.

As a Christian ethicist his biggest concern was the relationship between human beings and God, on one hand, and human beings to each other, to their communities and to the world, on the other hand. Torleiv Austad in a review of Niebuhr’s work characterized him as “one of the most influential theological ethicists of the twentieth century.”[176] What makes Niebuhr remarkable is that his “loyalty to Christ, gratitude to God and responsibility for the world” is still making him a “great inspiration for systematic theology, especially in the field of social ethics.”[177]

Niebuhr was a contemporary of Bonhoeffer, who happen to live on the other side of the Atlantic Ocean. Both Bonhoeffer and Niebuhr started to develop their ethics of responsibility around (at least) the same time, though “there was no interaction between the two (even though his brother Reinhold might have told him about his German student and friend).”[178] After World War II, Niebuhr continued his development of a systematic understanding of responsibility. For

[176] Torleiv Austad, “‘The Responsibility of the Church for Society’ and Other Essays by H. Richard Niebuhr,” *European Journal of Theology* 21, no. 1 (April 2012): 72.

[177] Austad, “‘The Responsibility of the Church for Society’ and Other Essays by H. Richard Niebuhr,” 72.

[178] Huber, “Ethics of Responsibility in a Theological Perspective,” 197.

Niebuhr, it is not "*responsible life" an sich*, as with Bonhoeffer, but the "*responsible self"* that is the focus of his research and publications.

Niebuhr developed his concept of responsibility in a series of papers and lectures in the post-World War II period. A (partial) synthesis of that work, based on the lectures he gave at the University of Glasgow, the Pacific School of Religion and Riverside Church, appeared posthumously in 1963 in the book *The Responsible Self*.[179]

Putting the responsible self at the center reveals Niebuhr's interest in an anthropological foundation of ethics, and more particularly in relational anthropology. A foundation that he further developed in dialogue with philosophy, social sciences, and the humanities. Niebuhr referred to the dialogical philosophy of Martin Buber's "I-Thou-relation," but even more to the work of the American philosopher George Herbert Mead. The step from "Buber to Mead is of specific importance because Niebuhr learned from Mead a sensitivity for not only the dual but the triple structure of the interpersonal relationship."[180] I will elaborate on this triple structure further on in Subsection IV.4.3.

Niebuhr develops his concept of responsibility to overcome what he sees as major shortcomings of the then-dominant ethical approaches. These were teleological ethics and deontological ethics, which Niebuhr links to in the metaphorical language he used in his book to describe the man-the-maker and the man-the-citizen image, respectively. The man-the-maker is "acting for an end, gives shape to things, of course, refined and criticized in the course of its long use, by idealists and utilitarians, hedonists and self-realizationists."[181] Such a man wants to see

[179] Niebuhr, *The Responsible Self: An Essay in Christian Moral Philosophy*.

[180] Huber, "Ethics of Responsibility in a Theological Perspective," 197.

[181] Niebuhr, *The Responsible Self: An Essay in Christian Moral Philosophy*, 49.

his life as purposeful and his actions are driven by questions such as "What is my goal, ideal, or telos?" followed by "What shall I do?"[182] Practically, this results in ends-and-means reasoning. This makes rules for man-the-maker "utilitarian in character" and the "[a]ll laws must justify themselves by the contribution they make to the attainment of a desired or desirable end."[183] The defect of the man-the-maker approach that in dealing with ourselves as persons or as communities, neither the end nor the means are under our control, is dealt with within the man-the-citizen approach by focusing on what is possible. When considering what is possible in the environment or community in which one lives, such a man must ask questions: "To what law shall I consent, against what law rebel? By what law or system of laws shall I govern myself and others? How shall I administer the domain of which I am the ruler or in which I participate in rule?"[184] For man-the-maker, responsibility is moving towards goals, and laws and rules should be utilitarian; they should be a means to ends. For man-the-citizen, responsibility is being moved by respect for laws, and the good is subjected to the right; there is no future ideal only a present demand.[185]

The above two ethical approaches, indicated by the synecdochic analogy of the maker and citizen image, though widely used may obscure relevant aspects of man's self-defining conduct. This is clear when a man is suffering. For then, humans experience their limits "in which the self defines itself by nature of its responses."[186] In these events, humans experience

[182] Niebuhr, 60.

[183] Niebuhr, 55.

[184] Niebuhr, 53.

[185] Niebuhr, 55.

[186] Niebuhr, 59.

that life is no longer under control and that another law than ours is intruding “our self-legislating existence.”[187] Based on the possible self-defining characteristic of the way a man responds to suffering, Niebuhr posits his alternative to the above two ethical models: the ethics of man-the-answerer.

In this man-the-answerer approach, the human person as a self in relationships is of special importance. A further presentation of this approach is given in two parts. The first part, which may be well understood from a secular perspective, is presented in the following three subsections. The first two subsections discuss the responsible self in terms of two concentric circles of increasing scope of responsibility:

1. The inner-circle discussing the responsibility of the individual self or community
2. The outer circle treats the responsibility of the individual self or community in the surrounding society.

After these two subsections, the time dependence of these circles is discussed.

The second part is theologically oriented and discusses the center or reference point of these concentric circles. This is done by first viewing what makes the *Responsible Self* unique and, second, how acting responsibly is related to, respectively, sin, guilt, and, salvation.

After the summary of Niebuhr’s work about *The Responsible Self*, two more topics are discussed. These topics relate to George Herbert Mead’s work on the social self and Niebuhr’s views on human conscience. They are discussed in the final two subsections.

[187] Niebuhr, 60.

IV.4.2 Responsibility of the Individual Self or Community

The example of suffering highlights that "[m]an's self-conduct begins with neither purposes nor laws, but with responses." Man starts with asking questions, such as "What is going on?" and "What is the *fitting* action?"[188] Starting from this example, Niebuhr builds up a theory of responsibility that attempts to see the self in a living relationship with other humans. Such relationships with others become distorted but divine grace and love can also renew them. Niebuhr's theory consists of four elements.

The *first* element is the idea of *response*. This is only the action of a self or moral action unless it is a response to an *interpreted* action that has been exerted on the self. This *second* element of interpretation distinguishes responsive actions from pure reflexes, like your eyelids reacting to light. This interpretive element is what characterizes human awareness that in a more or less intelligent manner "identifies, compares, analyzes, and relates events so they come to us, not as brute actions, but as understood and as having meaning."[189] When determining a response, we may not so much use laws as guides of our own conduct but rather to predict the response of the other we are reacting to or who will react to us. Interpretation is not only a rational matter of the consciousness, but it is also influenced by feelings of aggression, guilt, and fear that are invoked in the encounter of the other. These feelings may be buried in our deep memory that is only partly under our immediate control.

The third element is the *anticipation of a reaction to our reaction.* Next to the interpretation of actions upon the self, the responsible self accepts the consequences of the reaction of others and this does not prevent the self from continuing the interaction.

[188] Niebuhr, 60–61, italics original.

[189] Niebuhr, 61.

The final and *fourth* element is *social solidarity*. As one could not speak of a responsible self when its response originates from a source that is completely different from where the action came, the element of social solidarity assumes that we respond to action impinging upon us in "a continuing discourse or interaction among beings form a continuing society."[190] Based on these four elements, I now state Niebuhr's summary about responsibility:

> The idea or pattern of responsibility, then, may summarily and abstractly be defined as the idea of an agent's action as a response to an action upon him in accordance with his interpretation of the latter action and with his expectation of response to his response; and all of this is in a continuing community of agents.[191]

It appears that this response-action ethics of man-the-answerer does not fit biblical interpretation. For most interpreters of the Old Testament, and for theologians such as Barth and Bultmann, the ethics of the Bible is an ethics of obedience.[192] Contrary to that, or rather in addition to that, Niebuhr also views the Bible as a testimony of humans led by decisive questions, such as "What is happening?" and "What is the fitting response to what is happening?"[193] Therefore, Scripture can also be used to support the new way of thinking of Niebuhr.

IV.4.3 Responsibility in Society

Stimulated by the importance of the social understanding of human conscience, as explicated by the social psychologist George Herbert Mead, Niebuhr posits that the responsible self never responds in an isolated I-Thou interaction. In our encounters with others, we experience

[190] Niebuhr, 65.

[191] Niebuhr, 65.

[192] Niebuhr, 66.

[193] Niebuhr, 67.

approvals and disapprovals of these others. Those responses of others take place in a larger context that demonstrates continuity, generalizing the particularity of the other. This makes Niebuhr conclude that the self's conscience "represent not so much [its] awareness of the approvals and disapprovals of other individuals in isolation as of the ethos of [its] society, that is, of its mode of interpersonal interactions."[194] Niebuhr views this ethos of society as the present Third or third reality in the encounter of I and Thou, which is characteristic of his triadic interpersonal relationship structure. In this triadic structure, the self, for example, discovers natural events. There is a constancy in the interpretation of these events as they are categorized by prior information derived from social, historical reason by the self's companions. This third reality is always participating in the communication between social partners. That makes the self responsible "in the sense of accountability when the response is made not to one being alone but to that being related with the self to a third reality."[195]

In addition to natural events, where we use speculative or observing reason as binding elements in the I-Thou encounter, there may also be another third reality, the "idea of the cause," or the loyalty with which we approach actions and the events we experience. Then, for example, commitment to a cause may influence our response to actions. A third example of this third reality is the connection of the self to a reference group related to the present challenges the self is facing. Such a reference group may even be transcendental, referring beyond themselves, for example, a universal society or God.

[194] Niebuhr, 78–79.

[195] Niebuhr, 82.

In a Christian community, this third reality can be prophets and apostles who represent a common cause, but it may also be Jesus Christ who can refer beyond Himself to the Creator as a common cause. Niebuhr introduces and defines the notion of universal responsibility as follows:

> The notion [of] universal responsibility, that is, of a life of responses to actions which is always qualified by our interpretations of these actions as taking place in a universe, and by the further understanding that there will be a response to our actions by representatives of universal community . . . or by an impartial spectator who regards our actions form a universal point of view, whose impartiality is that of loyalty to the universal cause.[196]

IV.4.4 The Importance of the Aspect of Time and History in Responsibility

The temporal aspect has restrictive importance in teleology where the future may be taken into consideration by the potential that it bears or by the ideal goals it allows to forecast, while the past may refer to sin and guilt that one has to bear. In deontology, the temporal aspect is not relevant, as one has to deal with supra-temporal laws.

However, in the action-response analysis of Niebuhr, time is highly relevant as the past and the future of a self are extensions of its presence, and, therefore, both belong to it in the now. Therefore, "[p]ast, present, and future are dimensions of the active self's time-fullness."[197] In this time-fullness, the interpretations of actions upon the self are also made concerning the future, past, and present. When, in this extended timespan, the focus is only on the actual instant, defense ethics will most probably result. Then our current understanding of the world is filled with enemies next to some friends, classifying our world as either good or evil; things that ought to be and those that are not to be. This leads to a life of self-preservation. This life view easily extends toward predicting that others also act in self-defense, conforming to what they did in the

[196] Niebuhr, 88.

[197] Niebuhr, 93.

past. The question then arises of how a self in its time-fullness is free or can change its historical context. In other words, how can the self reinterpret the present action to develop a new reaction in the future so that it is no longer an action out of self-defense?

Niebuhr outlines two ways in which such adaptability can be obtained, and both depend on the nature of the encounter. The first occurs when man encounters natural events. Then adaptation can occur when one breaks with the traditions of the past. This process starts with radical doubt by which every notion and form is questioned (anew) under the present conditions. The attitude of radical doubt has been the basis for scientific inventions. It required meeting old patterns of interpretations with fresh ones and changing the responses that were inscribed in man's social memory. It is not a matter of being obedient or disobedient but of establishing the response that better fits future encounters with natural forces. The second method of adaptation is more amenable to change our responses to persons and to communities who cannot forget or leave behind "in every present an internal, remembered past", and it is done by "*reinterpreting* the past."[198] Studying the past that has brought us into our new present allows such reinterpretation by recalling, accepting, understanding, and reorganizing the past instead of abandoning it. For example, on a personal level, analytical psychology tries to reconstruct one's personal past. For Christians, the reinterpretation of the past can be done by "forgiveness of sin, the remembrance of the guilt and the acceptance of the acceptance by those against whom they had offended," can create new freedom in their present.[199]

Repentance and forgiveness contribute, to a large extent, to the hope for the future because our response to actions in our present is based on the predictions we make of the future

[198] Niebuhr, 102.

[199] Niebuhr, 104.

and our recollections of the past. Moreover, based on the confidence we have in these predictions and in the expectancy of other actions rather than our own, while trying to consider constraints due to future scarcity or abundance, we can adapt these future predictions. The free will in all this is the agent who commits himself "to resolute questioning of the adequacy of his stereotyped, established interpretations."[200]

"[T]hese social and personal reinterpretations of remembered pasts and anticipated futures, however, [do] not radically change either our general pattern of understanding of action upon us or our general mode of fitting response as long as our sense of the ultimate context remains unrevised."[201] This revision of what is a fitting response in a lifetime and a history surrounded by eternal life is everlasting work. This ultimate context is further explored in the next subsection.

IV.4.5 Theological Dimension

Niebuhr developed a better understanding of the responsible self that is theological, although without a Biblicism that purports to base itself solely on Biblical texts. The two ethical issues of the responsible self that are addressed in a theological setting are, first, the uniqueness of the self in absolute dependence and, second, the acting of the responsible self concerning sin, guilt, and salvation. In both these issues, reference to the One beyond all, the One creative power, the Transcendent One, which, in a monotheistic view, is called God, becomes relevant.

[200] Niebuhr, 106.

[201] Niebuhr, 106.

IV.4.5.1 The Absolute Dependency of the Self

In addressing the first issue, namely the absolute dependency of the self, two existential questions are addressed: first, “What is the radical act by which the self is unique?” and, second, “How is the self evaluated in all its interpretations and actions?”

According to Niebuhr, the answer to the first question is that the uniqueness is a consequence of the absolute dependence by which the self is established because the self cannot interpret the radical action by which it came into existence and by which it lives in the here and now. We can attempt to forget this or, as done primarily in the East, welcome it. Either of these two responses is primarily a matter of faith. ‘Faith’, in this sense, is not in terms of some set of beliefs to be accepted in place of knowledge, but it is an “attitude of the self in its existence toward all existences that surround it, as beings to be relied upon or to be suspected.”[202] Such an attitude enables a human being to trust or distrust itself and this makes faith an ingredient in all what we can know. This is because such faith does not contradict knowledge but may, in trust, generate new faith or, in distrust, generate new knowledge. “The response in trust or distrust to the radical act of the self‘s and its world‘s creation qualifies all particular interpretations of finite actions upon the self and therefore all its reactions.”[203] However, the question arises of whether the self does not become fragmented in all these different actions and different roles that it conducts. The answer to this question by Niebuhr is a radical ‘no’ and he explains this by reverting to religious language to posit that it is “the soul and God that belong together,” or “I am one within myself as I encounter the One in all [who] acts upon me.”[204] Of course, there are

[202] Niebuhr, 118.

[203] Niebuhr, 121.

[204] Niebuhr, 122.

social aspects in the responses of the self, such as the loyalty of its social circles. They may fail the self due to their finiteness, but faith as trust and distrust in God affect and are affected by all the interactions of trust and suspicion among the Thou's and It's. This is why Niebuhr can assert "In this faith, by this faith, I live."[205]

In answering the second question, Niebuhr does not turn to sociologists or psychologists, because sociologists doubt whether the same person always acts in different roles. For psychologists, the process of self-distinction is unclear. In answering the second question about responsibility, Niebuhr states, "by that action whereby I am I in all the roles I play, in reaction to all systems of action that imping on me, I am in the presence of the One beyond many. And my response to every particular takes the form of response to the One that is active in it."[206] Therefore, in essence, in all its actions and interpretations, the self is responsible to God, since God is acting in all actions upon the self and, therefore, the self is to respond to all actions upon it to respond to His action.

IV.4.5.2 Understanding the Self in Relation to Sin, Guilt, and Salvation

In addressing this second issue of the theological dimension, Niebuhr presents an alternative based on his response-analysis for overcoming the paradoxes that deontology and teleology introduce to understand this relationship.

According to deontology, the paradox centers on the problems of law and the gospel. This paradox arises when reconciling the requirements of the law, on one hand, with the love to God with all your heart, soul, mind, and strength, and the love to your neighbor as to yourself.

[205] Niebuhr, 120.

[206] Niebuhr, 122–23.

Similar to this paradox, another one arises “in the reflection that the action of the redeemed must be obedient to the will of another than the self, namely, God, and yet that if redeemed, it be done in freedom, namely, in the doing of one’s own will.”[207] This second paradox restrains human freedom to the obligation to obey God. Right life is being obedient to this and other rules and sin is the transgression of these rules. Salvation from sin and its consequences is the justification of the transgressor, requiring the condition of repentance “interpreted as [an] acknowledgment of guilt and sorrow for sin” and “perhaps also the substitutionary punishment of another, the Christ.”[208]

According to teleology, the paradox is its persistent direction in the pursuit of an ideal. A major challenge is to reconcile this pursuit with the Christian conviction and the primacy of the experience of God‘s action, in which He reveals His goodness. “Sin is understood as loss and confusion, rather than guilt.”[209] Loss arises by abandoning the pursuit of the ultimate good and replacing this pursuit by seeking lower good ends. The departure from pursuing this ultimate good causes confusion. Salvation then becomes restoring the goal or granting the self another chance to grow towards perfection. The required healing can be found in the power of the self that flows from seeing God, and from living in His likeness.

By focusing on the human experience of the self in dealing with sin, guilt, and salvation, it can experience both wretchedness and glory, bondage and freedom, death and life. In this response analysis, Niebuhr expands his image of the responsible self to overcome the paradoxes of the man-the-maker or man-the-citizen image by viewing God as a friend, rather than the

[207] Niebuhr, 131.

[208] Niebuhr, 130.

[209] Niebuhr, 133.

enemy, as One we can trust. This image calls for a new interpretation. Salvation now appears to us as deliverance from that deep distrust in God, a distrust that "causes us to interpret everything that happens to us as issuing ultimately from animosity or as happening in the realm of destruction."[210] The importance of this difference in perspective on our actions and those impinging on us is that "[r]edemption appears as the liberty to interpret in trust all that happens as contained within an intention and a total activity that includes death within the domain of life, that destroys only to re-establish and renew."[211] When restricted to Christians, this new interpretation only becomes possible when viewing the action of God in all in connection to the response by one man who accepted becoming human and answered his death with resurrection. In Jesus Christ, the responsible self sees an exemplary life of obedience to the law, but going beyond the law; of pursuing "action which is fitted into the context of universal, eternal, life-giving action by the One."[212]

This relationship with Christ allows the Self to understand its values and its life as a divine call to be sensitive to what is going on in the world around it.

IV.4.6 Relation to George Hebert Mead's work on the social self

Niebuhr developed his theological ethics in dialogue with the social sciences and the humanities, especially with philosophy.[213] This is exemplified by the fact that the background of his theory of responsibility was neither in the theology of Ernst Troeltsch, emphasizing historical relativism,

[210] Niebuhr, 142.

[211] Niebuhr, 142.

[212] Niebuhr, 145.

[213] Huber, "Ethics of Responsibility in a Theological Perspective," 199.

nor Karl Barth's neo-orthodoxy, insisting on the primacy of revelation. Rather, it was infused by the insights from George Herbert Mead that the subject is not constituted of "a self, giving laws to itself, judges itself, approves or condemns itself," but emerges from a process of interaction.[214]

The influence of Mead on Niebuhr's work on responsibility became clear long before the publication of *The Responsible Self*. For example, his publication on "The Ego-Alter Dialectic and the Conscience"[215] had already mentioned it. In that paper, Niebuhr builds upon Mead's works, *Self and Society* and *The Philosophy of the Present*, to present an alternative to Kant's idealism, on the one hand, and Westermarck empiricism, on the other hand.[216] Instead of viewing the self as a separation between reason and emotions, with the empiricists identifying, for example, the emotional factors with the social entourage of the self, social theorists such as Mead developed the fundamental insight that the self knows itself through the mediation of another. Moreover, in such relationships with 'the other,' the self is an integrated agent that is both rational and emotional; it can be either subject or object.[217]

Niebuhr, however, criticizes that view of social theorists by calling their interpretation of 'the other' singularly narrow, for, too often, they conceive the self as living in a single society. According to Niebuhr, the self does not live in one society but in many communities. Therefore, the self does not only deal with one "generalized other", as the social theorists claimed, according to Niebuhr, but with many others that are not all generalized. Niebuhr explains such

[214] H. Richard Niebuhr, "The Ego-Alter Dialectic and the Conscience," *Journal of Philosophy* 42, no. 13 (June 1945): 355, https://doi.org/10.2307/2018981.

[215] Niebuhr, "The Ego-Alter Dialectic and the Conscience."

[216] Niebuhr, 352.

[217] Niebuhr, 353.

generalized other as “a sort of composite photograph which the self makes of the associates in his society.”[218]

By this criticism, Niebuhr viewed the moral life in which the self participates as “one in which a reasoning and feeling self takes toward itself the attitude of another which it represents to itself or which is represented to it.”[219] He sees the other as able to participate in different societies. For example, the self’s family, represented by the parent that must be obeyed, or a professional representative who judges the self on his professional ethics, or in a Christian community, this other examines the self in the eyes of Jesus Christ, or the self being confronted to the most general other, or as the God of revelation.[220] This versatility of the other contributed essentially to Niebuhr’s triadic relationship to describe the self’s participation in developing its moral life. This triadic structure is discussed in Section IV.4.3.

The distinction between Niebuhr’s and Mead’s model of the self outlined above is a generally accepted view. However, Joshua Daniel recently contested this view in an essay entitled “H. Richard Niebuhr’s Reading of George Herbert Mead.”[221] To better understand the contribution of Niebuhr, I briefly summarize this criticism below.

According to Daniel, the singularity that Niebuhr assigned to Mead’s view of the other concerning the self’s moral development is based on a misunderstanding. According to Daniel, “Mead’s account of the I/me distinction within the self is rich enough to describe the self’s

[218] Niebuhr, 354.

[219] Niebuhr, 355.

[220] Niebuhr, 354.

[221] Joshua Daniel, “H. Richard Niebuhr’s Reading of George Herbert Mead: Correcting, Completing, and Looking Ahead,” *Journal of Religious Ethics* 44, no. 1 (February 18, 2016): 92–115, https://doi.org/10.1111/jore.12133.

constitution in multiple communities."[222] This observation of Daniel supports that within the self, there should be negotiations between the "I" of the self and its multiple "me's," each of the latter representing the self's constitution in a community. This multiplicity of the "me's" aligns the I/me distinction of Mead with the dialectic in Niebuhr's model of the self. In this model, these negotiations between the "I" and "me's" are interpreted as the vulnerability of the self to the moral community in which the self participates, but also, conversely, when interpreted as the vulnerability of the community to be transformed by their individual members.[223]

We could then question the novelty of Niebuhr's triadic model of the self. Daniel asserts that this novelty is in the theological completion that Niebuhr's model brought through his concept of faith.[224] This completion consisted of broadening the account of Mead's teacher, Josiah Royce, who viewed normal life as an affair of loyalty. Royce's concept of loyalty fits into Niebuhr's triadic model of the self: loyal self, loyal others, and common cause.[225] According to Niebuhr, this common cause is interpreted as *social faith* or trust in a particular community, from which the self, on the one hand, draws its worth and its standards of axiological perception but, on the other hand, to which the self is loyal. Such social faith explains nationalism. However, and this is where Niebuhr goes further than Royce, he also defines *radical faith*. "The loyalty of radical faith is expressed dually as loyalty to each particular existent in the community of being and to the universal community."[226] For Niebuhr, social faith is taking a particular community as

[222] Daniel, 100-1.

[223] Daniel, 103.

[224] Daniel, 105.

[225] Daniel, 105.

[226] H. Richard Niebuhr, *Radical Monotheism and Western Culture, With Supplementary Essays* (Louisville, KY: Westminster John Knox Press), 33–34..

the One beyond all particular communities, while radical faith enables one to identify the creator with the redeemer.[227]

IV.4.7 Niebuhr and the Human Conscience

For this thesis, I draw attention to Jeffrey Morgan's rediscovery of Niebuhr's work on human conscience.[228] This rediscovery is necessary, according to Morgan, as the topic of the conscience, which he defines as "an individual's moral awareness before a locus of accountability and judgment" has been neglected by Christian ethicists.[229]

The suspicion about the conscience among ethicists of the twentieth century – and I would also like to add the beginning of the twenty-first century here – is due to Immanuel Kant. In synthesizing Kant's work on the conscience, Lehman concludes that this work opens the door for humans "to act according to laws of our own making and to judge ourselves by evaluating the rightness and wrongness of these actions with a seemingly divinely sanctioned moral certainty in proportion to the inner depth of that conscience."[230] Such view promotes subjectivity and individualism, and it aptly describes a modern man who is a free, autonomous, moral agent whose authority is not derived from external reality.[231]

In an attempt to rescue the notion of conscience for Christian ethics, Morgan turns to Niebuhr's triadic model of the self. Recollecting this contribution here will enlighten that model

[227] Joshua Daniel, *Transforming Faith: Individual and Community in H. Richard Niebuhr* (Eugene, ORE: Pickwick Publications, 2015), 172.

[228] Jeffrey Morgan, "A Loss of Judgment: The Dismissal of the Judicial Conscience in Recent Christian Ethics," *Journal of Religious Ethics* 45, no. 3 (August 14, 2017): 539–61, https://doi.org/10.1111/jore.12189.

[229] Morgan, 539.

[230] Paul Lehmann, *Ethics in a Christian Context* (New York:, NY Harper & Row, 1963), 336.

[231] Morgan, "A Loss of Judgment," 545.

even further. Niebuhr considers the multiple communities in which the self participates as multiple consciences. These consciences may sometimes be at odds with one another, causing tension within the self. To deal with this tension, a responsible self "must make a kind of meta-ethical choice about which of these consciences it is going to endorse."[232] From a theological perspective, Niebuhr brings the aspect of freedom that the responsible self has to interpret its life according to its center of valuation, which provides the self with a sense of "an ultimate community of interaction relative to which the others are penultimate or spurious."[233] The ultimate community of interaction opens the possibility for the self to relate to God. That freedom allows the self in times of conflict with other communities to step back, turn inward, and discern the relation to God. In the warning given by Niebuhr at the end of the previous subsection, it is important not to reduce this relationship to God to that of one selected community (of faith). Morgan ends his article with an invitation to Christian ethicists to rekindle interest in the human conscience.

As human conscience is highly relevant when determining how humans deal with advanced technology, I return to this topic in Chapter VIII.

IV.5 Potential of the Ethics of Responsibility for ADS

This section briefly discusses the potential of the theory of responsibility formulated by Bonhoeffer and Niebuhr concerning the ethical challenges of ADS. Moreover, I outline what is necessary to concretely apply their frameworks in this context.

[232] Morgan, 549.

[233] Niebuhr, *The Responsible Self: An Essay in Christian Moral Philosophy*, 101.

Bonhoeffer, as we saw, offers a structure of the responsible life. He provides a framework or landscape defined by four vertices in which a responsible life is lived. Of direct interest in the context of ADS is the vertex of considering action in accordance with reality, in other words, responsibility needs to be directed to concrete neighbors in their concrete reality. This provides several generic constraints to consider when wanting to act responsibly. These constraints, however, require concretization to operate in the reality of embodied AI. For example, it is recommended that responsibility must 'have the courage to look into the immediate future'. While this is an interesting notion, how do we realize this in practical ADS with AI and modern hardware? Similar remarks can be made about the other actions in accordance with reality, as well as with the other vertices.

Niebuhr's focus on the self also decidedly speaks of *philosophical* anthropology. He presupposes that human beings intentionally grasp and shape reality by responding to actions exerted upon them, interpreting these actions, and anticipating the response to our response. This interpretative capability demonstrates that humans, in essence, are social beings; it challenges human beings to make moral choices. As interpretation looks both to the past and future, it becomes possible to create something new in the present. The fact that we are historically conditioned means, contrary to teleology, that there is no unconditioned reason or idea. Furthermore, contrary to deontology, we cannot be ruled by immutable principles or ideals. This, however, does not exclude that goals and principles can guide us and that we may consider them when determining our response to an action. However, these goals and principles cannot become a principle in itself.

The interpretative aspect entails the crucial ability of the self to test and judge its responses by itself and to evaluate that of others and society. Moreover, in this evaluation, the self is confronted with the "limits which order life and beyond which life is destroyed."[234] Sedwick gives the example that by breaking trust and loyalty, which happens in personal relationships and social life, we experience the suffering of the destruction of life.[235] Such destruction of life is sin and induces guilt. Instead of attempting to deny these symbols by redirecting one's goals or trying to mold laws by reinterpretation, James Gustafson proposed an ethics of redemption in his introduction to Niebuhr's book *The Responsible Self*. Here, God's action upon us in redeeming us from sin, evils, and death gives us the freedom to act responsibly. It redeems us from the need for self-justification. It motivates us to give freely because we freely receive; it motivates us to forgive because we are forgiven; it makes us grateful. "God's redeeming action ... is a transformation and transvaluation of all our actions."[236] This importance of living in God's grace is also crucial in Bonhoeffer's structure of responsibility.

The works of Bonhoeffer and Niebuhr, reviewed in this chapter, provide frameworks to understand the moral life and building blocks for a normative stance on ethics, in which reality and the human self are central to the determination of (ethical) responses as opposed to teleological and deontological ethics. They, however, do not provide form or content for addressing concrete moral questions. Further concretization is necessary in both cases. For the topic of this thesis, it is necessary to concretize an ethics of responsibility to address ethical questions of ADS. Then, concrete ways of reasoning about responses, interpretation, dealing with limits and constraints to bring knowledge into practice are necessary. Two approaches of

[234] Timothy F Sedgwick, "Niebuhr's Ethic of Responsibility: A Unified Interpretation," *Saint Luke's Journal of Theology* 23, no. 4 (September 1980): 281.

[235] Sedgwick, 281.

[236] James W. Gustafson, *Introduction of The Responsible Self: An Essay in Christian Moral Philosophy* (New York, NY: Harper & Row, 1963), 39–40.

concretization are analyzed in the remainder of this thesis. First, in Chapter V, the programming of ethical decision-making in ADS, which was considered in a top-down approach in Chapter III, is again taken up but now in a bottom-up manner. This considers the development of an approach based on the work of Thornton[237] to program ADS ethically. This approach anticipates a clear role of the designer (programmer) of ADS. Second, from Chapter VI onward, I consider the second sub-question introduced in Chapter I and focus on human responsibility in the design and operational use of ADS. Here, it will be shown that while the major tasks of driving can be taken over by ADS, human involvement in key decisions remains necessary. In this analysis of the human responsibility in ADS, the theories of Bonhoeffer, but especially Niebuhr, will show their usefulness. They allow one to analyze the shortcomings of existing design frameworks to avoid responsibility gaps in the design and the use of ADS, to propose new improved solutions based on that analysis.

[237] Sarah Marie Thornton, "Autonomous Vehicle Motion Planning with Ethical Considerations" (PhD. diss, Stanford University, 2018), https://searchworks.stanford.edu/view/12746436.

Chapter V

Programming AI for ADS Ethically: A Bottom-up Approach

V.1 Introduction

In Chapter III, I demonstrated that the ‘classical’ ethical approach to study the integration of ethical decision-making in autonomous vehicles is to apply various normative ethical theories, such as ethics of deontology or teleology, to hypothetical life-threatening traffic dilemmas, such as the trolley problem. This approach is unsuccessful in programming ethics in AI for ADS. The main reasons are the lack of realism of the abstracted hypothetical traffic scenarios and the individualistic focus. The lack of realism is also a consequence of the search for a single or set of rules to make a single split-second decision. Bonhoeffer and Niebuhr’s approach to responsibility is to consider the real situation. For example, the work of Bonhoeffer about acting in accordance with reality provides clear guidelines for ethics to consider reality. As I will reference Niebuhr when addressing the second subquestion of this thesis in Chapters VII and VIII, this chapter mainly uses Bonhoeffer’s approach to review several bottom-up ethical decision-making strategies for ADS. The guidelines of Bonhoeffer, translated in an ADS context, are the following:

- consider the actual and relevant driving conditions
- consider the limitations of ADS’ ability to act
- the goals of a good outcome and good intentions need to be quantified and realized in reality
- attempt to understand the entire given reality
- the action should be based on weighting, evaluating, observing the given situation while considering the limits of human understanding

— the action should be taken while considering the immediate future and the consequences of these possible actions

— the action should consider the actions of the other stakeholders

If the ADS decision-making considers the above suggestions and constraints continuously or regularly until a positive outcome is achieved, it will contribute to the realism of the ADS decision making. The limited period to act is a direct consequence of the limitation of the relevance of a certain act. For example, when overtaking a car, the relevance of the act is over when the maneuver is completed.

Control engineering has the potential for algorithmically determining actions in ADS with realism. Especially modern control engineering methods that make use of mathematical models of the object, such as of the car to be controlled while driving over an icy road, can consider the actual and relevant driving conditions. Model-based control methodologies include *Model Predictive Control* (MPC)[238] and *decision-making under uncertainty*.[239] Control engineering aims to determine an action based on observations, which is similar to that of many AI methods. Moreover, modern control engineering methods are based on the use of optimization methods that further strengthen the parallel with AI. In the scope of this thesis, I consider the use of control engineering methods in ADS as one particular way to integrate AI in ADS. A mathematical model is essential to introduce further realism into the action determination. As MPC uses the mathematical model to estimate the effect of an action in the short-term future, it becomes possible to anticipate the future impact of control actions, such as

[238] J. B. Rawlings, “Tutorial Overview of Model Predictive Control,” *IEEE Control Systems Magazine* 20, no. 3 (June 2000): 38–52, https://doi.org/10.1109/37.845037.

[239] Mykel J. Kochenderfer et al., *Decision Making Under Uncertainty: Theory and Application*, Illustrated edition (Cambridge, MA: The MIT Press, 2015).

the short-term effect on the car's behavior as the consequence of braking. Further, a mathematical model allows, within the limits of that model, to understand the entire reality that is relevant to a particular traffic scenario, such as whether the tires will start slipping. Therefore, modern control engineering methods can be used to automate the control activities at the tactical and operational level in Michon's driver model (Section II.2). The main problem from an ethical perspective is that the objective, in other words, the anticipated good outcome of the action, is often based on engineering performance specifications. For example, when the top tactical level of Michon's driver model has planned a particular trajectory of the car to follow from start to end, a performance specification could be to follow this trajectory as closely as possible. The latter is then expressed by a reward function that penalizes the deviation of the actual car trajectory from the specified one.

The reward functions generally lack direct ethical interpretation, which hampers the ethical interpretation of these modern control-engineering methods, and, hence, prevents control engineering methods to take ethics into consideration. To allow the control-engineering expert to make justified ethical choices, it is necessary to understand the 'behind' of these ethical choices as well as the trade-offs that are possible between different human values. One should not merely look for a particular set of rules that are valid for a single traffic momentary snapshot situation, as was considered in the top-down approach in Chapter III. This requires the development of a systematic framework in which normative ethical theories can be integrated into the control engineering design, and that justified trade-offs between the different normative ethical goals and control goals can be made. In such a framework, these goals should not become principles. Therefore, the goal is not to program ethics in ADS but to program ethically, as envisioned by philosopher Valor.[240] Vallor presented this vision at the workshop "Towards Programming Ethics in Automated Vehicles" held at Stanford University in June 2015 to challenge developers of AI not to try to program ethics in machines, but to develop their character as developers so that they program ethically.[241] In this type of framework, the control engineer should be better able to make ethical trade-offs, but human freedom and character are still necessary to make the final selection. It is still a matter of debate who must make the final selection: the individual driver, the manufacturers, or society through legal norms.

This chapter aims to highlight the initial steps taken in the field of control engineering to develop such a systematic framework for programming ethically. This is accomplished in two steps. First, I present – via an example of a concrete realistic traffic scenario – how ethical considerations can systematically be integrated into the control engineering design. The integration of the normative deontological and consequential ethics, engineering performance specification, as well as considering and resolving conflicts between them are presented. This is done in Section V.2 based on the work of Thornton et al.[242] The integration is achieved mathematically by the consideration of multi-criteria optimization. The potential of programming ethically is further shown by demonstrating its extension to two relevant ethical

[240] Shannon Vallor is a Professor at Santa Clara University in California. She is an AI ethicist who focuses on the impact of emerging technology and science on the human character. One of her major works is Shannon Vallor, *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting* (Oxford, New York, NY: Oxford University Press, 2016). In this book, Vallor challenges humans to cultivate their moral character by developing what she calls 'technomoral virtues' such as 'technomoral honesty'. The latter virtue is respect for truth by expressing truth in appropriate ways, for example, by considering the sensitivity associated with various media channels. This statement is also made in Thornton, "Autonomous Vehicle Motion Planning with Ethical Considerations," 61.

[241] Gereon Meyer and Sven Beiker, eds., *Road Vehicle Automation 6*, Lecture Notes in Mobility (New York, NY: Springer International Publishing, 2019), 44, https://doi.org/10.1007/978-3-030-22933-7.

[242] Sarah M. Thornton et al., "Incorporating Ethical Considerations into Automated Vehicle Control," *IEEE Transactions on Intelligent Transportation Systems* 18, no. 6 (June 2017): 1429–39, https://doi.org/10.1109/TITS.2016.2609339.

aspects. The first extension is the additional consideration of human values in the control-engineering framework, which is done in Section V.3.1. The second extension is the explicit focus on the 'other', which was the central incentive in Bonhoeffer's vicarious representative action. This focus on the other in automated driving aims for a 'fair' (or just) distribution of driving opportunities among different traffic participants in mundane traffic scenarios. Section V.3.2 briefly considers different ethical notions of what fairness means. The chapter concludes with some remarks in Section V.4.

V.2 Incorporating normative ethical theories in automated vehicle control

To keep the outline focused on the ethical issues, I present this incorporation via a realistic case study. This is an adapted version of the scenario that was considered by Thornton,[243] adapted to Highway Code of Belgium. I will start sketching this adapted scenario in the following subsection.

V.2.1 Traffic Scenario

An automated vehicle is driving on a two-lane roadway at a constant speed. An obstacle obstructs the lane of that vehicle, and that obstacle is still far enough ahead of the vehicle to make different choices. A full white line separates the two lanes of opposite traffic. According to Article 72.2 of the Belgian Highway Code, it is prohibited to cross such a white line.

The first option is to give priority to continue moving and enter the adjacent lane to move around the obstacle. This would cause the vehicle to briefly violate the traffic law. It is common

[243] Thornton et al., 1430.

practice that human drivers cross a full line, for example, when overtaking a cyclist.[244] This seems to indicate that in some cases like crossing a full white line "human compliance with traffic laws seems less deontological" and inferior to mobility.[245]

The second option is to sacrifice mobility and strictly adhere to the traffic law to not cross a full white line.

The challenge this traffic scenario posits for the programmer of the ADS is to be able to make trade-offs between conflicting ethical aspects as well as engineering goals and allow for continuous ethical decision making instead of an isolated (rule-based) ethical decision like was considered in the top-down approach. Such a systematic integrative approach in the design of automatic controllers for ADS becomes possible when the approach enables the translation of ethical requirements in terms of control engineering reward functions and/or constraints, and when the approach allows mathematically treating the trade-offs. Therefore, I discuss in the next subsection the specification of different ethically relevant objectives as reward functions that can be treated next to control engineering reward functions in the same mathematical framework.

V.2.2 Specification of Reward Function

The different specification and their control engineering interpretation are:

1. *Trajectory Tracking*: What is called a "good" outcome is that the *tracking error*[246] should be as small as possible. Usually, a sum of squared values of this tracking error, like in

[244] Thornton et al., 1433.

[245] Thornton et al., 1433.

[246] The tracking error is the difference between the predicted trajectory of the vehicle over the (short-term) future timespan considered by the MPC controller and the desired trajectory, as specified by the strategic part of the ADS. The future timespan of the MPC controller is a time window of fixed length that moves ahead of the actual time instance the control action is executed in a receding manner.

regression analysis, is taken as a reward function. Similarly, other engineering performance specifications, such as limiting the side accelerations of the vehicle, can be specified by a similar reward function as the sum of squared values of (static) transformations of the full state of the vehicle.[247]

2. *Obstacle Avoidance*: "Obstacle avoidance is a high priority in navigating roadways . . . and the choice of assigning collision avoidance as a deontological rule is natural."[248] Such a rule could be the geometric constraints that determine the allowable space the vehicle can move in safely. Such space is typically described by a virtual tunnel that is mapped on the route of the vehicle (along the receding future time-horizon presented in item 1). For a fixed virtual tunnel, this would result in hard constraints. For example, for the second choice, it would be the lane in front of the vehicle until the obstacle. However, to allow for the first choice, Thornton opted to make these virtual tunnel constraints soft, via the so-called slack variables in optimization theory.[249] This softening of the constraints turns the overall control problem into a minimization problem of a reward function. The solution of this minimization problem would allow the selection of a tunnel that bi-passes the obstacle safely and thereby makes the ADS to cross the full white line. By the use of slack variables, it becomes possible to violate the constraint (slightly) thereby ultimately preserving safety (that is in this case avoiding the obstacle).

[247] The full state of the vehicle are the physical quantities, like the vehicle's displacement, velocity, angular rates, etc. that fully characterize the behavior of the vehicle. All these state quantities, in the generic case, uniquely represent what has happened in the past to be able to predict the future.

[248] Thornton, "Autonomous Vehicle Motion Planning with Ethical Considerations," 43–44.

[249] Thornton et al., "Incorporating Ethical Considerations Into Automated Vehicle Control," 1433.

3. Traffic laws: Traffic laws generally give rise to hard constraints to the state components of the vehicle. A typical example is restrictions on the speed limit of vehicles. These are deontological by nature, formulated in terms of structure and rules. However, to allow for the possibility to (slightly) violate them, as often done by human drivers, the control engineer has the option to either consider them as hard constraints or soft constraints, as discussed in item 2, via the introduction of related slack variables.

The specification in terms of the sum of squared errors or slack variables can be seen as consequentialist (utilitarian) specifications. This is because the optimization attempts to minimize them. Therefore, minimization may be considered as trying to make the outcome as close as possible to the stipulated goals in real-time. On the other hand, the rules in terms of limits on the state are deontological by nature. The introduction of slack variables allows turning deontological constraints into consequentialist ones. To have a unified systematic framework, a weighting parameter has been introduced in the framework of Thornton et al, that when weighting the slack variable with a 'large' value, the corresponding constraint becomes more deontological while for a small value it becomes more consequentialist.

V.2.3 Fine-tuning the Controller

Based on the specifications discussed in the previous subsection, the control design problem becomes a constrained weighted minimization problem that is subjected to the mathematical model of the car. A single reward function is optimized that is a weighted sum of the individual reward functions considered in Subsection V.2.2, whereby the individual terms are either a sum of squared quantities derived from the state or slack variables. The inequality constraints are (some of) the deontological constraints (not translated to consequential rewards via the use of slack variables). When the weights in the single reward function are chosen, the overall design becomes a constrained optimization problem. This optimization approach then allows understanding what the limits are of the design. The understanding of these limits contributes to acting in accordance with reality, as explicated by Bonhoeffer.

Therefore, a key element in the fine-tuning of the systematic methodology is the selecting of the weights. As outlined in the previous subsection, this weight selection allows making a constraint either deontological or consequentialist. This depends on the choice and, hence, on the freedom of the human being making this selection.

The human involvement thus remains necessary in the fine-tuning of the automatic controller. This freedom can be given to, for example, the owner of the car, the control design engineer, or the legislator prescribing the weights. This indicates that, though the systematic approach of Thornton allows putting normative ethical choices in a control engineering framework and bringing this in relation to the actual conditions, the final choice still requires human intervention. It is here that the other two vertices of Bonhoeffer's structure of responsible life become crucial. They motivate the individual to act in his or her own freedom, not hiding behind a rule and because of that freedom to be willing to "embrace guilt." An example is the trespassing a full white line may be the ADS operated car to not obstruct or to further complicate the overtaking of the obstacle by an emergency vehicle. The concern is then focused on the 'other' human being to be rescued. This would require that the human operator of the ADS be able to make the selection of the weight before the operation of ADS. This is again a clear indication of the necessity of human involvement in ADS.

V.3 Extensions

There are two important extensions to further integrate normative ethical theoretical aspects in the controller design for ADS. The first is the integration of human values and the second is a just distribution of driving opportunities to traffic participants. Both extensions are briefly discussed subsequently in the following two subsections.

V.3.1 Human Values in the Design of ADS

Next to the ability to make a trade-off between deontology and consequentialism in ethical decision-making, the consideration of human values should play an immanent role. It is one of the cornerstones in the *Matrix of Christian Ethics*, where a human value is defined "not so much as *thought* as it is *experienced* within a particular context."[250] Value indicates what human(s) consider as important in a particular situation they are involved in. The methodology sketched in the previous section enables us to consider the values of safety, efficiency, and comfort of the human occupants of the vehicle. However, an extension is needed to consider values that are more relevant for more stakeholders involved. For example, when a vehicle operating under ADS is approaching a pedestrian crossing "the direct stakeholders are now the occupants in the automated vehicle, the pedestrian potentially crossing the street, and the authority of traffic laws."[251] The values of interest in the context of an ADS approaching a cross-walk, are "mobility, safety, legality, care and respect for others, fairness and reciprocity, respect for authority, trust and transparency and individual autonomy."[252] These are defined as (taken from Thornton et al.[253]):

— *Care and respect for others*: This is manifested by the desire to avoid harming others.

[250] Nullens and Michener, *The Matrix of Christian Ethics*, 55, italics original.

[251] Sarah M. Thornton et al., "Toward Closing the Loop on Human Values," *IEEE Transactions on Intelligent Vehicles* 4, no. 3 (September 2019): 439, https://doi.org/10.1109/TIV.2019.2919471.

[252] Thornton et al., 439.

[253] Thornton et al., 439.

— *Fairness and reciprocity*: Both affect the occupants of the vehicle as well as other stakeholders (pedestrians, etc.) in that the automated vehicle should not take biased or discriminatory action based on information about the stakeholders. The automated vehicle should treat all stakeholders equally.

— *Respect for authority:* This engages the relationship between the automated vehicle and its adherence to traffic laws.

— *Trust and Transparency:* Trust emerges when the pedestrian assumes an oncoming vehicle yields to his or her right-of-way while crossing the crosswalk. Transparency occurs when the automated vehicle's action facilitates trust.

— *Individual Autonomy:* This is the autonomy of the vehicle's occupants acknowledging the desire to get from one destination to another with little impedance.

Thornton et al. have developed a Value Sensitive Design methodology to translate these values into a control engineering design framework.[254] This required two fundamental extensions of the approach outlined in the previous section. First, it is necessary to make the controller design methodology able to deal with uncertainty in the values. For example, when a pedestrian is to develop trust in the ADS, that system should be able to cope with the probabilistic nature of whether or not a pedestrian approaching a crosswalk will actually cross it or not. For that purpose, use can be made of the knowledge that "the likelihood of the pedestrian transitioning from the sidewalk to the crosswalk is a function of the pedestrian posture."[255] Considering this probabilistic nature in a decision-making problem can be done using the methods described by

[254] Thornton et al., 437-46.

[255] Thornton et al., 440.

Kochenderfer et al.[256] Second, these human values have to be translated into reward functions that can be considered in the control optimization step. This has been fully documented for the above list of values in the work of Thornton et al.[257] Again a multi-criteria optimization framework adding weighted variants of all the reward functions is proposed to derive the best fitting solution. The weights now can be used as design parameters "to determine conflicts between stakeholders and values."[258] The way such conflicts can be handled is by an iterative design procedure where each iteration consists of a conceptual, technical, and empirical step. The conceptual step entails the definition of the stakeholders and the human values reward functions; the technical step consists of multi-criteria optimization and the empirical step "allows for quantitative and qualitative analysis of the developed design, such as data analysis or observations from human-user studies," where the actual control design is tested in realistic simulation studies and actual real-life trials.

Such iteration process requires human judgment to change the weights, modify stakeholders and model description to make the overall design better fitting with the human experiences. This is a second illustration of the relevance of human involvement in the design for ADS. Bonhoeffer's structure of responsibility can again shed useful light in this environment. For example, the vertex of vicarious representative action would invite the AI designer to think and reflect on the other traffic participants by taken a stance in their place. This might lead to conflicts between the designer and the company policy. Such conflicts might be similar, though

[256] Kochenderfer et al., *Decision Making Under Uncertainty*.

[257] Thornton et al., "Toward Closing the Loop on Human Values."

[258] Thornton et al., 438.

less breath-taking, to the conflicts Bonhoeffer faced between acting responsibly or hiding behind rules.

V.3.2 ‘Just’ Distribution of Driving Opportunities

When considering multiple vehicles either with or without ADS or other traffic participants, such as bicyclists, pedestrians, emergency vehicles, etc. the first vertex of Bonhoeffer’s structure of responsibility becomes a crucial issue. One of the central issues in taking care of ‘the other’ in traffic with ADS is how to program an ADS such that it justly (or fairly) distributes a limited traffic space among other traffic users of the same space. This topic of just distribution of traffic opportunities in mundane traffic situations has been considered by Dietrich and Weisswange.[259]

When ADS is to make decisions by taking all other relevant traffic participants in the particular traffic scenario into account, ADS must act “so that driving opportunities of the traffic participants in the scene are justly distributed.”[260] To make this approach explicit two steps need to be taken. The first is “to find ways to quantify driving opportunities, determined by factors such as safety, utility, and comfort.”[261] These values can (again) be considered as reward functions for control engineering design as was illustrated in the previous section, but now such reward functions are defined for all relevant traffic participants. This requires for a single ADS the estimation of the state of these other participants. The second is the definition of the principles according to which the distribution of these opportunities is considered just. Different

[259] Manuel Dietrich and Thomas H. Weisswange, “Distributive Justice as an Ethical Principle for Autonomous Vehicle Behavior beyond Hazard Scenarios,” *Ethics and Information Technology* 21, no. 3 (September 1, 2019): 227–39, https://doi.org/10.1007/s10676-019-09504-3.

[260] Dietrich and Weisswange, 232.

[261] Dietrich and Weisswange, 232.

ethical principles could be used for that purpose. Examples are the utilitarian principle or the "difference principle of Rawls' theory."[262] The use of both principles is discussed briefly in the remaining part of this subsection. For the implementation of these two steps, I refer to the paper of Dietrich and Weisswange.[263]

V.3.2.1 Distributing Justice Using Utilitarian Principles

Applying the utilitarian principle to *justly* distribute the driving opportunities to different participating traffic participants in a particular scene can be done by "choosing the option for the ADS for which the overall driving opportunities of all traffic participants are in sum the highest. In the implementation of this policy, the driving opportunities have to be expressed as reward functions to be derived from estimated trajectories of the traffic participants by the ADS.[264] The utilitarian principle would then select that policy that maximizes the sum of the reward functions of all traffic participants.

The critique of the application of this principle is inherent to the utilitarian approach. Namely, it does not prevent discrimination of a small group of persons. An example being that "options are chosen for which the majority receive a high utility but the minority is confronted with a relatively dangerous driving situation."[265]

[262] John Rawls, *A Theory of Justice*, rev. ed. (Cambridge, MA: Harvard University Press, 2009), https://www.hup.harvard.edu/catalog.php?isbn=9780674000780.

[263] Dietrich and Weisswange, "Distributive Justice as an Ethical Principle for Autonomous Vehicle Behavior beyond Hazard Scenarios," 234–37.

[264] Dietrich and Weisswange, 234–37.

[265] Dietrich and Weisswange, 233.

V.3.2.2 Distributing Justice using Rawls' Difference Principle

The objective of Rawls' difference principle is to make the prospect of the vehicle with the lowest expected driving prospects among all vehicles in the scene as high as possible. This principle is subordinate to the principles of equal liberty and equality opportunity, which Rawls sees as the leading principle. In implementing a scenario based on the difference principle, again the reward functions of all traffic participants, expressing their driving opportunities, have to be predicted by the ADS. In case quantities are not available in the calculations worst-case approximations can be used.[266] The difference principle would then select that policy that maximizes the minimal cost or minimal prospect.

For example, consider the following particular traffic scenario. A vehicle operating under ADS wants to overtake a slower driving truck in front of it, but before it starts the necessary lane change it 'sees' a fast-approaching car appearing behind it in the lane it needs to do this overtaking. For this case, if the cost for braking of the fast-approaching car "is very high, the car with ADS is requested to brake to let this fast car pass first."[267]

V.4 Concluding Remarks

This chapter has shown that the field of mathematical optimization as applied in AI and control engineering is making progress to enable the program ADS ethically. This means that trade-offs can be made among various normative ethical theoretical frameworks in the design of the control or intelligence of the ADS.

[266] Dietrich and Weisswange, 235.

[267] Dietrich and Weisswange, 237.

In the presented multi-criteria optimization methods, the ethical trade-offs require the tuning of the weights of the different rewards functions, each representing particular engineering performance specifications and normative ethical values, goals, and constraints. The selection of these weights still needs to be done by human experts during the design of the ADS but may also require human driver intervention before the operation of the ADS. This tuning highlights the freedom the human designer and/or operator still have in the overall decision for a particular choice. This use of this freedom can be based on Bonhoeffer's 'free venturing of a concrete decision', as outlined in Section IV.3.6. It may motivate a human driver to opt for violating deontological rules, like crossing a white line, when priority is given, for example, to not further obstruct an emergency vehicle.

This chapter has also shown the possibility of bringing ethics and engineering together from the onset of the design of the engineering system, in this case, ADS. This is not only the case for the mapping of ethical norms and criteria to the control engineering domain but also for interpreting the empirical validation studies, such as conducted in the iterative third step of the VSD approach in Section V.3.1. Such integration is still in its infancy as both communities have so far mainly been operating distinctly from each other. This has resulted in a "gap between rapid technology growth and slower ethical reflection on the consequences of that growth". [268] The contributions presented in this chapter will decrease this gap as it calls for collaboration between ethical experts and designers of the ADS intelligence during the design phase.

Finally, this chapter has shown that this new integrative approach still foresees an important role for the human being both during the design as well as before the operation of the

[268] Nullens and Michener, *The Matrix of Christian Ethics*, 31.

ADS. In the following part of the thesis, I will attempt to demonstrate that the human operator (driver) needs to be actively involved during the operation of ADS, though the latter may do most of the driving.

Chapter VI
Meaningful Human Control over ADS

VI.1 Introduction

In the previous chapters, the focus has been on answering the first sub-question of this thesis formulated in Chapter I. This first sub-question was about integrating normative ethics into the design of ADS. However, an important ethical challenge remains, namely the question of how to deal with responsibility during the operation of ADS. In this chapter, as well as in the following ones, the second sub-question of the thesis will be analyzed.

ADS, like many autonomous machines, performs tasks and functions that are ordinarily performed by humans. An example is the automatic lane changing to overtake a vehicle in front of the car with ADS. However, such automated systems can during their operation experience failures by which they select and engage inappropriately. This can cause serious and vast military, humanitarian, and ecological disasters. An example was the misclassification of a truck as a part of the sky in the Tesla S example reported in Example 2 on page 133. Even though such autonomous systems are developed, manufactured, tested, and deployed by humans, the autonomous machine itself selected the inappropriate target. This stipulates the challenge, 'who is accountable for this disaster?' Is it the human operator denying frequent warnings of the systems, the engineer programming the ADS, or the autonomous machine? According to Michael Horowitz, Adjunct Senior Fellow of the US Technology and National Security Program, machines cannot be moral agents;[269] as a result, attempts are made to give 'more' authority back

[269] Horowitz and Scharre, "Meaningful Human Control in Weapon Systems: A Primer."

to the humans involved in the operation of automated systems. Such an attempt is summarized under the heading of "meaningful human control" (MHC).

Initially, this phrase was introduced in Article 36 of a report of the United Kingdom about the use of autonomous weapon systems.[270] This article did not present an additional requirement for the design of future autonomous weapon systems but rather presented the principle of MHC to ensure that these systems are used in compliance with the laws of war. However, what the article failed to do was to clarify what the notion 'meaningful' means. Despite attempts to further clarify this term, such as in the working paper of Horowitz and Scharre,[271] the difficulty in defining the term is highlighted in the following statement: "[P]olicy-makers and technical designers lack a detailed theory of what "meaningful human control" exactly means and therefore they do not know which specific legal regulations and design guidelines should be derived from this principle."[272]

An attempt to lay such a theoretical basis was made by Santoni de Sio and van den Hoven.[273] In this chapter, I first summarize this theory and then apply it to ADS. In Section VII.2 of the next chapter, I bring this application in dialogue with Niebuhr's ethical response-action model of *The Responsible Self*. This will result, first, in highlighting several shortcomings. Second, and more importantly, it will motivate the definition of a new theoretical framework that

[270] United States Department of Defense, Autonomy in Weapon Systems, 3000.09. UK Government, "Article 36, Killer Robots: UK Government Policy on Fully Automated Weapons," April 2013, http://www.article36.org/wp-content/uploads/2013/04/Policy_Paper1.pdf.

[271] Horowitz and Scharre, "Meaningful Human Control in Weapon Systems."

[272] Kerstin Vignard, "The Weaponization of Increasingly Autonomous Technologies: Considering How Meaningful Human Control Might Move Discussion Forward," *UNIDIR*, 2014, https://www.unidir.org/files/publications/pdfs/considering-how-meaningful-human-control-might-move-the-discussion-forward-en-615.pdf.

[273] Santoni de Sio and van den Hoven, "Meaningful Human Control over Autonomous Systems," 1-14.

I have called *Responsible Human Control over ADS*. This theory is outlined in Chapter VII. The first steps in concretizing this new theory for ADS are presented in Chapter VIII.

VI.2 Fundamental Ideas of the Theory of Santoni de Sio and van den Hoven

The idea of the theory is to let the adjectives 'meaningful' and 'human' promote a stronger and clearer connection between the automated system and the human agents involved in the design and operation of that system. This idea should lead to increased safety and clearer accountability. The theory should "not only accommodate all relevant moral considerations, but also be suitable to give ethical guidance to policy makers, engineers, and technical designers.[274] In this way, the theory is an extension of the Value Sensitive Design method outlined in Section V.3.1 and allows the analysis of the accountability of all humans involved in the design and operation.

The developed theory aims to overcome the following flaws still often taken to heuristically cope with AI in human-operated systems:

- The human presence is not sufficient for being in control of a system, with control here understood in its general sense of having the possibility to influence its outcome according to a desired goal with the system.[275] This is because a human operator may not be able to causally influence crucial parts of the system. For example, when the AI system overrules the human operator (driver of an ADS vehicle) in a life-critical traffic scenario, or one may not have enough information to appreciate what is going on, a similar situation is created like

[274] Santoni de Sio and van den Hoven, 3.

[275] Later on, an ADS specific definition of the notion of *Control* will be given in Section VII.3.4. See the definition on page 115, taken from Sven Nyholm. This definition states that *Control* is the ability to supervise and have full authority over the agent under control.

when “a human being is instructed to push a button when a light bulb goes on without knowing that in fact by doing this an attack missile is launched.”[276]

— Respecting the user manual is not sufficient for being in control. This may often contain misleading or lacking information about the performances of the AI system. An example is the selling argument of the Tesla S model promoted as an ‘autopilot’ system, giving the driver the impression that he or she does not have to look after the vehicle during scenarios the vehicle is in this auto-pilot mode. See Example 2 on page 133.

— In the use of autonomous weapon systems, Mary Ellen O’Connell[277] and Asaro[278] advocate that when initiating an individual attack the meaningful human control conditions are guaranteed by insisting on the presence of a human operator to take a “near-time decision.” However, it is unclear what near-time in this sense means. Furthermore, when extending this condition to ADS, it would require the human driver to be alert nearly ‘most’ of the time without understanding the need for it.

VI.3 Philosophical Foundation

The foundational question in moral responsibility according to Santoni and van den Hoven is “whether and under which conditions humans are in control of and therefore responsible for their

[276] Horowitz and Scharre, “Meaningful Human Control in Weapon Systems,” 10.

[277] Mary Ellen O’Connell, “Chapter 12. Banning Autonomous Killing: The Legal and Ethical Requirement That Humans Make Near-Time Lethal Decisions,” in *The American Way of Bombing Changing Ethical and Legal Norms, from Flying Fortresses to Drones*, ed. by M. Evangelista and H. Shue (Ithaca, NY: Cornell University Press, 2018), 224–36, https://doi.org/10.7591/9780801454578-014.

[278] Peter Asaro, “On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making,” *International Review of the Red Cross* 94, no. 886 (June 2012): 687–709, https://doi.org/10.1017/S1816383112000768.

everyday actions."[279] In addressing this question, a compatibilist approach is taken. Following present-day compatibilists, such as Frankfurt,[280] Dennett[281] and Fischer, and Ravizza,[282] they reject that "any delegation of decision-making to non-human agents amounts *per se* to a disappearance of human moral responsibility over decisions and actions."[283] The proposition for such a claim is often made to a priori abandon autonomous weapon systems, because of the reasoning that the control of a machine over human life of death is morally wrong, a position that is also taken by the Holy See.[284]

Based on the theory of Fischer and Ravizza, a person is morally responsible for an action X if that person possesses "guidance and control over that action."[285] Under this guidance and control law, the person opting for action X should use a decisional mechanism that is (1) "moderately reason-responsive", and (2) the person is the owner of that decision mechanism ("ownership").[286]

Fischer and Ravizza primarily focus on a decision mechanism related to one human being. For humans, condition (1) of guidance and control requires that agent (human) to act

[279] Santoni de Sio and van den Hoven, "Meaningful Human Control over Autonomous Systems," 4.

[280] Harry G. Frankfurt, "Freedom of the Will and the Concept of a Person," *Journal of Philosophy* 68, no. 1 (January 1971): 5–20, https://doi.org/10.2307/2024717.

[281] Daniel C. Dennett, *Elbow Room: The Varieties of Free Will Worth Wanting*, NED-New edition (Cambridge: The MIT Press, 2015), http://www.jstor.org/stable/j.ctt17kk7ns.

[282] John Martin Fischer and Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility* (Cambridge: Cambridge University Press, 1998).

[283] Santoni de Sio and van den Hoven, "Meaningful Human Control over Autonomous Systems," 5.

[284] Santoni de Sio and van den Hoven, 2.

[285] Santoni de Sio and van den Hoven, 5.

[286] Michael E. Bratman, "Fischer and Ravizza on Moral Responsibility and History," *Philosophy and Phenomenological Research* 61, no. 2 (2000): 453, https://doi.org/ppr2000612106.

according to a decisional mechanism that (1i) can recognize the presence of strong reasons to act (or not to act) and (1ii) bring himself to perform (or not) that action in a sufficiently broad range of circumstances.[287]

The conditions (1i-1ii) rule out the responsibility for actions taken "under excusing factors such as (non-culpably) being under the influence of potent drugs, direct manipulation of the brain, behavior attributable to a significant brain lesion or a neurological disorder, phobias, drug addiction, and coercive threats."[288]

Santoni and Van den Hoven are making two further remarks. First, the possessing of a decision mechanism of "actual dispositional features,"[289] is to be interpreted by the fact that the possibility for defining alternative scenarios for action does not mean that the agent himself has to bring about such alternatives. Second the focus of the theory of 'Guidance and Control' is not on the circumstances or motivational factors that the agent can manipulate, but rather on the characteristics of processes or 'mechanisms' leading to action, on their sensibility, flexibility, or lack thereof.[290]

The second condition for guidance and control (2) refers to the fact that the person acting in a morally responsible way takes ownership of the decisions taken. This requires (2i) the agent to perceive the effects of the decision taken on the world, (2ii) the agent to take into consideration that others may have moral reactions toward how the decision affects the world, and (2iii) considers that the requirements (2i-2ii) are to be based on the "agent's evidence in an

[287] Santoni de Sio and van den Hoven, "Meaningful Human Control over Autonomous Systems," 5.

[288] Fischer and Ravizza, *Responsibility and Control*, 35–36.

[289] Fischer and Ravizza, 52–53.

[290] Fischer and Ravizza, 38.

appropriate way."[291] To illustrate this notion 'in an appropriate way', Santoni and van den Hoven present the example of making decisions by tossing a coin. Then one should be aware that such decisions are based on statistics and, therefore, statistics is not to blame for the outcome or that others will hold you accountable for the decision.[292] Ownership is excluded in the case of "psychological manipulation, subliminal persuasion, strong nudging, strong entrapment, brainwashing, and indoctrination."[293]

Santoni and van den Hoven extend the theory of Guidance and Control of Fischer and Ravizza to consider multiple agents in the decision process and they allow agents to be artifacts or engineering systems. In this extension, "autonomous systems are part of the decision-making mechanism" and this opens the possibility of using Fisher and Ravizza's conditions for guidance and control.[294]

The extension by Santoni and van den Hoven translates the two conditions of Guidance and Control into a tracking respectively tracing conditions, now no longer exclusively related to one (human) agent but holding for a complete system of humans and non-human artifacts. The system perspective for ADS includes "the human agents (driver, designers) and vehicles, as well as the whole traffic environment and the social, legal, and political infrastructures."[295] The system perspective could be further expanded to not only include the operational staff of the latter environments and infrastructures but also the designers of it. Furthermore, such extension

[291] Fischer and Ravizza, 207–39.

[292] Santoni de Sio and van den Hoven, "Meaningful Human Control over Autonomous Systems," 6–7.

[293] Santoni de Sio and van den Hoven, 6.

[294] Santoni de Sio and van den Hoven, 6.

[295] Mecacci and Santoni de Sio, "Meaningful Human Control as Reason-Responsiveness," 106.

may also consider the educational and training system to improve the understanding, designing, and operating of autonomous systems.[296]

> The tracking condition requires a system to be responsive to relevant human reasons to act or to refrain from acting. The tracing condition requires instead the presence of one or more human agents in the system design history or use context who can at the same time appreciate the capabilities of the system and their own responsibility for the system‘s behavior.[297]

These two conditions characterize in general MHC, but the next two sections expand on their explanation for ADS.

VI.4 The Tracking Condition for MHC over ADS

When formulating the concretization of the tracking condition, I will consider both a design as well as an operational perspective.

The concretization starts with making an inventory of "all" the tasks involved in the operation of ADS and then tries to link these tasks to intentions or reasons of one or more humans who can execute each task. To make an inventory of the tasks for ADS, Mecacci and Santoni make use of John Michon‘s classification of tasks a driver (must) perform(s). These tasks are grouped into three functional levels, strategic planning, tactical maneuvering, and operational control. I refer to Section II.2 for more details. The hierarchy that exists in these functional levels is that the top level of strategic planning coordinates and constrains the lower level tasks of tactical maneuvering, which on its turn coordinates and constrains the lowest level of operational control.

[296] Santoni de Sio and van den Hoven, "Meaningful Human Control over Autonomous Systems," 12.

[297] Mecacci and Santoni de Sio, "Meaningful Human Control as Reason-Responsiveness," 104.

For Michon, each functional level of a vehicle system is under the control of an agent if its behavior responds respectively to the agent's strategic plans, tactical maneuvers, or operational tasks. Therefore, this notion of control could be applied to subsystems of an ADS. For example, the operational tasks could be fully executed by software according to what has been coordinated and planned by the human driver executing the strategic planning and tactical maneuvering tasks.

However, Michon's definition of control, stipulated in the previous paragraph, becomes problematic when different actors are involved in all levels of controls and when it is no longer clear which of these actors takes responsibility. This lack of clarity introduces the so-called responsibility gap as introduced in Chapter I. The topic of the responsibility gap is further discussed in Section VII.3.4 with the help of Niebuhr‘s view on the continuity in Social Solidarity. Mecacci and Santoni, in an attempt to resolve these possible responsibility gaps, demand that each element with control capabilities in the system, including the human agents themselves, should be *maximally responsive to reasons.* According to Mecacci and Santonio, this condition implies that all elements of the complete system should be able to act like humans who are capable, for example, by appropriate training or skills, to behave according to certain reasons.[298]

These other elements (for control) should not only be made responsive in the given sense but could "in turn offer relevant reasons for other components to recognize and to respond to."[299] By the extended system view given in Section VI.3, "an automated driving system should not only appropriately respond to plans of its driver but also to some relevant features of road

[298] Mecacci and Santoni de Sio, 106.

[299] Mecacci and Santoni de Sio, 106.

infrastructure — such as signs, traffic lights — as well as to some formal and informal traffic norms present in society."[300] Including these norms in the framework of analysis, opens up the possibility to consider the rules and norms as reasons or intentions of designers, policy makers, or the society in which they are embedded.

Mecacci and Santoni de Sio suggest the concretization of the tracking condition through the following three steps:[301]

1. The identification of which reasons are relevant for which humans in a given context.
2. The identified reasons of step 1 are categorized and ordered using a scale or metric, and an inventory is made of which part of the system can be designed to respond to each reason.
3. The definition of a *control map* that shows (3i) how the different reasons are interrelated and (3ii) how each reason is related to the overall system behavior.

In realizing these three steps normative decisions have to be made to make the control map 'objective' in order " to identify and prioritize different reasons of different agents."[302] As this normative aspect is user-dependent or dependent on a group of users, Mecacci and Santoni de Sio omit it from the design process of MHC for ADS. In the modeling of the control map in step three of the realization of the tracking condition, the theory of action of the analytic philosopher Elizabeth Anscombe is used, extended with the work of Stanford Professor Michael Bratman, to explain the relationship between actions, intensions, and plans of an agent, as

[300] Mecacci and Santoni de Sio, 106.

[301] Mecacci and Santoni de Sio, 107.

[302] Mecacci and Santoni de Sio, 107.

summarized in *Two faces of Intention*.[303] From this theory, a proximity scale is devised to realize the categorization and order in step 2 of the above three-step concretization. This proximity scale determines the 'distance' of each reason to the influence it has on the system's behavior. Figure 1 illustrates the effect of the use of this scale and the 'spatial' organization of the reasons in a map.

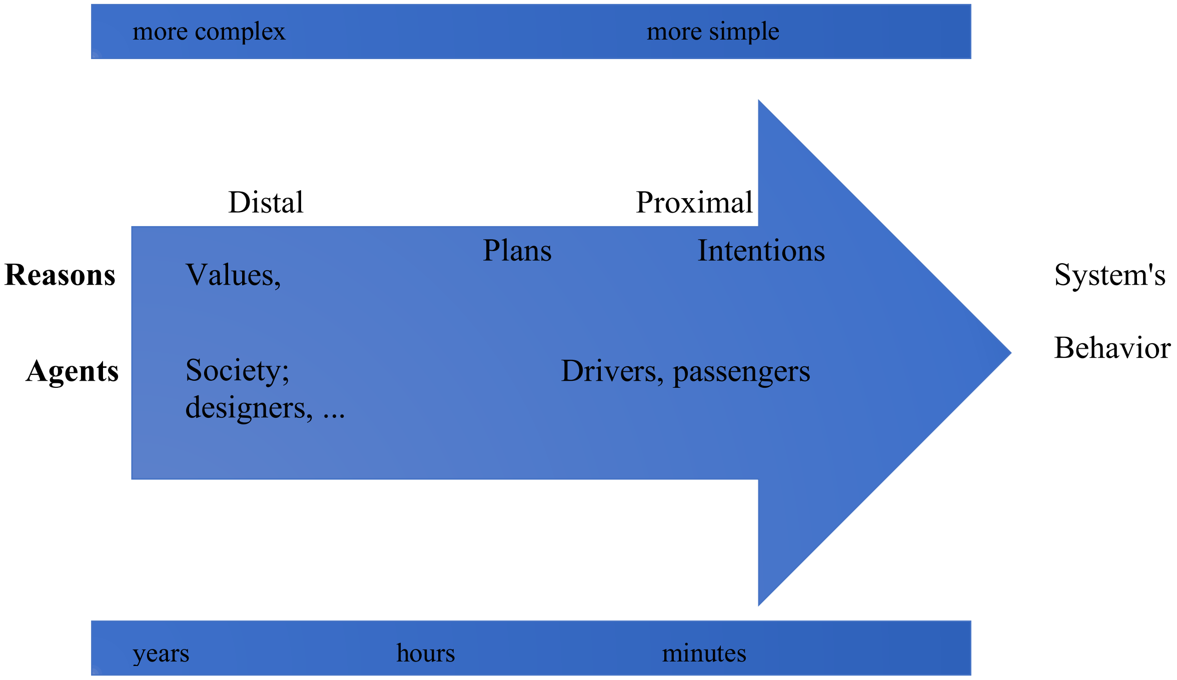

Figure 1: Visualization of the proximity scale in a control map. Reasons can be classified according to their proximity value. Bratman's proximal and distal intentions (plans) are typically temporally closer to a system's behavior. They are also simpler in the sense that more complex reasons explain and affect a system's behavior only through more proximal ones. Different agents can also be identified as typical endorsers of certain kinds of reasons.[304]

[303] Michael Bratman, "Two Faces of Intention," *Philosophical Review* 93, no. 3 (July 1984): 375–405, https://doi.org/10.2307/2184542.

[304] Mecacci and Santoni de Sio, "Meaningful Human Control as Reason-Responsiveness," fig. 3 and its caption.

Figure 1 depicts both time and complexity of the reasons as scales or metrics to determine the distance to the (actual) system's behavior. In this instance, the time factor is a relative time distance that separates two or more reasons and not the absolute time.

While this proximity scale allows one to determine the relevance of the reasons, it also helps to identify which agent typically plays a role related to the reasons. Typically, drivers or end users are linked at a more proximal distance to the system's behavior, while those governments coordinating traffic laws and regulations are linked to more distal reasons.

Example 2: Dual-mode ADS is a level 3 ADS following the SAE terminology defined in Chapter II. While such level 3 ADS only provides "partial driving autonomy in certain circumstances, such as while driving on a highway,[305] it was promoted by Tesla as an 'Auto-pilot' driving system. For such level 3 ADS, the driver is obliged to "constantly remain vigilant and ready to regain operational (manual) control at *any* time."[306]

In this auto-pilot mode, an accident was caused with a Tesla-S model in 2016 when the AI in the ADS misclassified a white truck in front of the car as a piece of the sky.[307] This caused a collision between the Tesla-S car and the truck, killing the driver.

[305] Mecacci and Santoni de Sio, 111.

[306] Mecacci and Santoni de Sio, 111.

[307] Danny Yardon and Dan Tynan, "Tesla Driver Dies in First Fatal Crash While Using Autopilot Mode," *Guardian*, June 30, 2016, sec. Technology, http://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk.

The use of this control map is illustrated for dual-mode ADS, defined in Section II.3. The realization of the tracking condition for dual-mode ADS starts with the first step to "assess the extent to which the semi-automated car is responsive to the relevant reasons of the relevant agents."[308] This assessment requires the answering of the question "whether and to what extent the system's behavior is responsive to the driver's relevant reasons to act?"[309] The human driver in the case of the Tesla-S accident had the following two basic reasons according to Mecacci and Santoni. First, he did not intend to take over the operational control task of the vehicle. This could be concluded from the fact that the driver did not touch the controls (steering wheel, pedals) for a long time before the fatal crash. Second, he intended to reach his destination safely.

The Tesla-S system violated the tracking condition, as it was not designed to respond to the driver's second intention. This is the consequence, first, of the fact of the misclassification error, but second, and more importantly, the Tesla-S system was not able to perceive that the driver did not and was not planning to intervene.

Example 2 and the above discussion following the example illustrate a responsibility gap between agents involved in the system. A solution, according to Mecacci and Santoni, is to re-design the system to fulfill the tracking condition. They present their solution in terms of the following simple two-rule-based Algorithm.

> **Rule 1:** Respond to a proximal reason if and only if it does not conflict with a more distal reason.
> **Rule 2:** Respond to the most proximal reason allowed by Rule 1.

[308] Mecacci and Santoni de Sio, "Meaningful Human Control as Reason-Responsiveness," 111.

[309] Mecacci and Santoni de Sio, 111.

In general, this two-rule algorithm tends to give priority to the driver's safety above individual freedom and flexibility. This algorithm further illustrates how moral reasoning might be integrated into the design of MHC.

Example 2 (Ct'd): In the application of the above two-rule-based algorithm to the Tesla-S accident, the proximal reasons of the ADS were the necessary driving maneuvers to advance according to the traffic plan and taking the traffic on the highway into consideration. The distal reason was to drive safely. The latter required the ADS to constantly check the driver's readiness to regain manual control. A condition the Tesla-S ADS failed to do.

This viewpoint makes the ADS master over the driver by constantly checking whether he is ready to take over control. While it is far from trivial how to do this checking and with what frequency, it does not consider the responsibility of the human operator.

VI.5 The Tracing Condition of MHC over ADS

The tracing condition of MHC "can be achieved only if it is possible to identify one or more human agents within the design and use a chain that has a capacity to (i) understand the capabilities of the system whole at the same time (ii) appreciate their own moral responsibility for its behavior."[310]

For the dual-mode ADS considered in Example 2, tracing is realized by appreciating the real limits of the ADS and that of the driver, as well as the fact that the appreciation of moral responsibility for (mis)behavior of the system is owed by the same (group of) persons. This tracing condition is realized, for example, when "the car manufacturers are well aware of the

[310] Mecacci and Santoni de Sio, 105.

technical limits of the driving systems they produce and/or of the (current) mental limits of a human driver for whom it is produced."[311]

One of the reasons the accident highlighted in Example 2 happened, was that the car manufacturer shifted all responsibility for the accident to the driver, by having them accept certain terms and conditions in the contract. The driver on the other hand did not know the full limits of the ADS and was not vigilant during the auto-pilot model of that ADS because of the 'confidence' in the ADS.

VI.6 Concluding remark

The tracking and tracing conditions for MHC over ADS might indeed improve the distribution of responsibilities among the different agents in the design and operation of the system.

The main driver in the design philosophy of the tracking and tracing condition is teleological ethics of reasons, goals, and intentions. Therefore, in the following Chapter VII, I will start with evaluating MHC over ADS using Niebuhr's man-the-maker model. In addition, deontological ethics in the form of rules and laws could be taken into consideration, but to appreciate the need for innovation I restrict in the next chapter to look at the man-the-maker facet.

[311] Mecacci and Santoni de Sio, 106.

Chapter VII
Responsible Human Control over ADS

VII.1 Introduction

The fundamental philosophy of MHC over ADS, see Section VI.3, is in essence based on the search for the conditions by which human drivers of a car operating under ADS should act, or if they do not, the ADS should take appropriate action. In principle, it is a search for those conditions by which humans are enabled to act in their freedom.

This fundamental search has been a topic of debate and continues to be so, among many scholars, such as philosophers, psychologists, neuroscientists, etc. It has resulted in various schools of thought, such as those of the incompatibilists and the compatibilists. Section VI.3 presented a brief description of the latter school of thought. The key hypothesis here is that humans make their decision based on an internal mechanism and the scientific objective then becomes one of tracing the nature, features, and operation of that mechanism.

Contrary to this reductionist view of human beings, Bonhoeffer and Niebuhr take a holistic view of human nature and its relationship to its Creator and His Son Jesus Christ. Hereby, Bonhoeffer in particular simply "eliminates the spurious question of determinism or indeterminism — whether the essence of the human spirit is to be falsely subsumed under the law of cause and effect."[312] Instead, for Bonhoeffer, responsible humans act in their own freedom, "without the support of people, conditions, or principles, but nevertheless considering all existing circumstances related to people, general conditions, or principles."[313] As proof of human freedom in their actions, Bonhoeffer sees the possibility that is granted to humans, "to

[312] Bonhoeffer, *Ethics*, 283.

[313] Bonhoeffer, 283.

observe, judge, weigh, decide, and act on their own, examining their motives, prospects, values, and meanings they want to express with the actions."[314] In the deliberation of their actions, humans cannot hide behind concepts like the purity of motives, favorable conditions, or meaningfulness of intended actions.[315] The only 'binding' element for such a 'free' human being is the bond to God and neighbor as they encounter the responsible self in Jesus Christ. That bond is liberating.

In addition, Niebuhr does not chase after some internal mechanism by which humans make their decisions. He arrives with his ethical model of response-to-action of a responsible self, at the insight that when one is looking for an "arbitrary free will" of a responsible self and its location, it can only be located "at the point where the agent commits himself to inquiry into further, longer series of interactions and into the responses taking place in a larger society, or at the point where he commits himself to resolute questioning of the adequacy of his stereotyped established interpretations."[316]

According to Niebuhr, the critical mind of the human being that investigates the facts and actions around himself is the source of acting responsibly. Such investigations are guided by the confidence we expect to have in our interpretations of the relevant actions and determine our responses. This confidence building in interactions with nature and others is crucial for human decision-making. It does not happen overnight but is inherently characterized by a learning process by which we use our human capability to revise our predictions of the way we and our environment will behave in the future to determine our present actions. This adaptive nature, as

[314] Bonhoeffer, 283.

[315] Bonhoeffer, 283.

[316] Niebuhr, *The Responsible Self: An Essay in Christian Moral Philosophy*, 106.

was highlighted in Section IV.4.4, is essential to acting responsibly. It does not mean always doing the right thing or acting according to the best ideals, but always trying to do the best by providing the most fitting response given the actual circumstances as perceived by the responsible self. This also makes the problem of dealing with failures critical. In dealing with this problem, both Bonhoeffer and Niebuhr take a Christological viewpoint. For Niebuhr, for example, failures of humans, which in a Christian-Judaic theology are sins, should not be tried to forget, by trying to forget the past. Forgetting the past could be devastating, as it is such a rich source by which we learn how to deal (in a better way) with the future. Therefore, humans should not forget their past failures, "but by forgiveness of sins, the remembrance of their guilt, and the acceptance of their acceptance by those against whom they had offended," that humans are set free to respond differently (better) in the future.[317] One appropriate way to learn to improve from past errors is by reinterpretation, as presented by Niebuhr in Section IV.4.4.

Based on the above exposure, I will develop an alternative ethical model for enabling a responsible interaction between human agents and ADS that stimulates collaboration and makes use of the learning capabilities of both the used AI and that of the human agents involved. The model should enable the analysis as well as the enhancement of responsibility in humans both in the development of ADS as well as during the operation of these systems. While I restrict mainly to the analysis of the level 3 ADS, as illustrated by the dual-model ADS in Example 2, the model might be useful to deal with human responsibility when operating with a much wider class of autonomous systems, such as medical robots. The new model, devised by the author of this

[317] Niebuhr, 104.

thesis, is indicated by “Responsible Human Control over ADS” to emphasize the human responsibility in control of ADS, though ADS may contribute to most of the driving.

The development of this new model is done in the following sections. In Section VII.2, the model of MHC over ADS of Santoni and van den Hoven is evaluated using Niebuhr’s ethics of the man-the-maker. Such evaluation allows highlighting the shortcomings of the model, but also allows us to fully make use of the solutions offered by Niebuhr in designing a new model. This Responsible Human Control (RHC) over ADS model and is presented in Section VII.3. This is done based on the four key elements of Niebuhr’s ethics of the responsible self, namely (1) response in the present, (2) Interpretation, (3) accountability, and (4) Social Solidarity. In the scope of ADS, attention is dedicated to two additional elements, namely (5) Adaptability and (6) Dealing with personal faults. Section VII.4 presents some concluding remarks. The RHC over the ADS model is concretized in the next Chapter VIII, which mainly focuses on dual-mode ADS systems.

VII.2 Evaluating the MHC over ADS of Chapter VI using Niebuhr’s man-the-maker ethics

The model to be evaluated is that of Santoni and van den Hoven, later on specialized to ADS system by Mecacci and Santoni in terms of Tracking and Tracing conditions. This method will be referred to in this section by the abbreviation “MHC-T&T”.

The man-the-maker ethics has its roots in Aristotle’s ethics. For Aristotle, “man is the being who makes himself — though not so by himself — for the sake of a desired end.”[318] Several elements of this idea could be related to core ideas of MHC-T&T. First the man’s activity to make himself could be a motivation for looking for a ‘mechanism’ by which

[318] Niebuhr, 49.

(internally) decisions are made. Second, this popular model for decision-making is ruled by human desires. In the same way, the MHC-T&T model is principally based on the desires, reasons/intentions of the agents involved. Thirdly, MHC-T&T attempts to link each reason or intention to at least one human agent. The advantage of this individualistic approach is that one can develop a hierarchy of reasons linked to possible actors, as done using the tracking and tracing parts of the MHC-T&T model, presented in Section VI.4. Such methodology tries to systematically collect the answers of all the agents involved to the central primary question of the man-the-maker model, namely "What is my goal, ideal, telos?[319] With this inventory of a control map, the tracing condition tries to navigate through the control map to allocate ownership by relating actors to reasons. Following the man-the-maker model, such ownership as well as the action, is determined by responding to the answer "What shall I do?"[320]

The major problem of the control map is that the information by which it is designed or by which users should operate is hidden. When restricted to the ADS only, leaving the driver out, it may be ok to present the control map in this way. However, when expecting the user the respond, he or she lacks information. For example, when what is suggested, due to the lack of information does not correspond with his or her expertise based on past images, a conflict may arise. To avoid such conflicts in addition to the what-question, ADS should also respond to how it arrived at this 'suggestion' or 'order'. For learning also addressing the why-questions are then highly relevant. The control map contains precise information and excludes confidence information, for example, in terms of probabilities related to both actions and actors. Including such information is not all straightforward. This topic is also not discussed in the MHC-T&T

[319] Niebuhr, 60.

[320] Niebuhr, 60.

approach. Furthermore, the discrete and fixed nature of the map in terms of a finite number of pre-defined reasons-actions-actors does not allow the human operator to learn from, neither to teach, the ADS. Even when a reason-action-actor connection has been identified in the control map, its effectuation may be very inefficient. For example, when a driver is not responding to a call from the ADS to show signs of vigor, should it stop the car immediately (at a safe spot), or should it wait for some time (how long?) to re-issue an alarm? Raising such alarms frequently without an actual need for the driver to interfere leads to so-called alarm fatigue. This creates a dangerous situation, as was the case in the Tesla-S accident of Example 2, that the driver might not respond when needed. This lack of robustness, in this case, is also a consequence of not presenting information to the driver by which he knows when to be alert and when not. Finally, the MHC-T&T approach may guarantee safety at the cost of significantly reduced performance and efficiency, as the above example for checking the driver's vigor has illustrated.

VII.3 The elements of the Responsible Human Control over ADS

Using the man-the-answerer ethics of Niebuhr's Responsible self, I will outline the new Responsible Human Control over ADS (RHC-ADS) approach. This outline is based on the first four elements of Niebuhr's definition of this man-the-answer ethics, namely:

1. Response in the present
2. Interpretation
3. Accountability
4. Social Solidarity

In addition, I will add the two elements:

5. Adaptability
6. Dealing with personal faults

These six elements will be described in the next subsections. Here I relate the insights from Niebuhr's responsible self to the responsibility of the human(s) involved in ADS. For simplicity, I will restrict ourselves to the human driver and the designer of the ADS only, leaving the extension to more parties for further research. The following chapter then will use these six elements to sketch a concretized solution towards distributing responsibility in (a simplified two agent) ADS system.

VII.3.1 Response in the Present

Instead of using reasons or intentions to distribute actions and the actor's responsibility, the Man-The-Answerer ethics' first element in the responsibility of the self is its *response* to an action. This can be a self-action or a reaction to actions of companions or others, but in all cases, the self and its action are always in the present. "I and now belong together."[321] It is this present that extends into the past and future of the acting self. The responsibility "lies in the agent who stays with his action, who accepts the consequences in the form of reactions and looks forward in the present deed to the continued interaction."[322] This posits the present as crucial in the definition of an agent's responsibility, ultimately linking the present deed to consequences from the past as well as to the potential the future offers.

The agent in the above statement about responsibility could be a human. However, within the scope of this report, the human who designed an artifact or computer program might act as a mediator between the world and another human. When the action-in-response is then taken concerning continuous time, the response could be viewed in a control systems engineering

[321] Niebuhr, *The Responsible Self: An Essay in Christian Moral Philosophy*, 93.

[322] Niebuhr, 64.

framework, discussed in Chapter V, as the response of a continuous-time controller[323] to measurements taken from its environment. For example, it could be a controller that measures the speed of the car and in response sets the gas throttle to keep that speed within a pre-specified interval. According to the above notion of responsibility, the designer who refrains to accept the consequences when the designed controller fails to operate according to specifications acts irresponsibly. When taking reference in time events, instead of continuous time, an event-based controller could then generate the response. This could be a human operator who responds to a control light turning amber or green. Such an event-based controller is typically used in alarming the pilot of an airplane when a failure of a system on board the aircraft, for example, one of the engines fails the button marked "ENG FAIL" will start flashing. The pilot is then triggered to respond to handle this failure. Proper training of the pilot is crucial to enable the pilot to handle such failures, next to many other things.

For responsibility, the present time or present event has to be seen at least concerning the response taken. Such response can depend on many other things, to be considered in the following subsections starting with the interpretations necessary to respond responsibly.

VII.3.2 Interpretation

"[F]or the ethics of responsibility the *fitting* action, the one that fits into a total interaction as response and as anticipation of further response, is alone conducive to the good and alone is right."[324] Crucial in determining such fitting action is the interpretation that starts with asking the

[323] Although in many modern engineering systems, use is made of digital computers as control systems, for the sake of simplicity in this report I only refer to analog controllers. This is because of the maturity of digital control theory by which the human operator perceives as if the digital controller is acting in a continuous manner. Therefore, I will refer to such analog or digital controller in this report as a *time-based controller*.

[324] Niebuhr, *The Responsible Self: An Essay in Christian Moral Philosophy*, 61.

question: “What is going on?” or “What is being done to me?” before the question “What shall I do?”[325] These questions combined with the intelligence of the agent allow the agent to develop understanding and discover the meaning of events happening to it through analyzing, comparing, relating, and identifying these events. When a gauge or a meter characterizes the event, interpretation is the discovery of the meaning the indicated value of the gauge has. For example, when a speedometer indicates a speed above the allowed speed one interpretation is that one is driving too fast.

However, often interpretation cannot simply be based on inspecting signals, like the value of a gauge, to our display. This is, for example, the case in interpreting the relations between managers and employees. Niebuhr suggests that a better understanding of such relations occurs when one grasps the way they are responding to each other’s actions according to their interpretations. This insight, according to Niebuhr, also holds between individual agents. Moreover, such a relationship ‘catalog’ could be indicated in engineering by a mathematical model. It allows predicting the response of the evolution of quantities of the system characterized by such model when certain actuator positions are chosen. For example, when having a model of a car driving on an icy road, the model allows predicting the pathway of the car when the car has a certain speed and the driver is trying to make a full stop. The act of prediction is an act of interpretation.

Our “interpretation equipment,” as if Niebuhr is thinking about a mechanism or even an algorithm that we are using in our interpretative activity, “binds us to [the] past.”[326] Our interpretations are “remembered images”, a way of emphasizing that it is stored in our memory,

[325] Niebuhr, 63.

[326] Niebuhr, 96.

and this ‘information’ is the product both of what society has taught us and to a lesser extent, to our past experiences in our encounters with other selves.[327] Though these images are based on past experiences, such as collected during training, we refer both to the future as well as to the past in interpreting present actions. The past gives us the images to deal with the future. This view to look ahead, maybe only for a short time window, corresponds with Bonhoeffer’s guideline in dealing with reality to have the courage to look into the immediate future as discussed in Section IV.3.4. Though Niebuhr also considers the larger context of our own life and our society in making our interpretations, he has in general a very conservative view on the way of anticipating the future. The conservatism stems from the evaluation scheme that humans use in classifying a situation or someone as being either good or evil, or things that ought to be or that ought not to be.[328] Therefore, Niebuhr’s responsible ethics is a kind of worst-case scenario that he says can explain, to a large extent, our actual ethics as defense ethics or an ethics of survival.[329]

An important element to reduce the conservatism of such worst-case reasoning is the confidence the self has in the actions in the past (of himself and others involved) to determine its response appropriately.[330] As our actions are based on the images of the past and what we have been taught in the past, the confidence in our actions of the past can directly be related to the confidence we have in these past images and lessons. To be able to assess this confidence is a crucial element in the collaboration between people. For example, when using Niebuhr’s worst-

[327] Niebuhr, 96.

[328] Niebuhr, 99.

[329] Niebuhr, 98.

[330] Niebuhr, 105.

case scenario we can be skeptical to collaborate with humans classified as evil or very reluctant to accept things that ought not to be trusted. Niebuhr's conservatism in interpretation can of course be relaxed by introducing degradations of trust or confidence that we have in humans or things, but trust or distrust and its degradations remain crucial in acting responsibly. An example of such degradations was given by the likelihood of a pedestrian crossing a crosswalk in Section V.3.1.

Humans have tried to develop intelligent machines that very much can imitate the above sketch of Niebuhr's view on interpretation. The following example illustrates this.

Example 3— Classification with Neural Networks (NN): I will not explain how NN works, for more details I refer to the paper of Fujiyoshi, et al,[331] and the references therein. In the scope of this thesis, I outline broadly what such a NN can do in relation to ADS. One important element for ADS is the classification of objects and their contours in a picture recorded with on-board cameras. Such classification is in essence addressing the 'What is going on?'-question and for example tries to determine whether the recorded picture contains a human and where this human, if present, is located. Imagine a picture of a human, perhaps an elderly woman pushing a trolley, in the middle of the street in front of the car. In determining which objects the NN detects in this picture, it compares the actual detected objects in that picture with pictures recorded in the past with the same camera of many different objects, such as different humans, animals, street lamps, garbage bins, etc. These are in essence fulfilling the roles of remembered images stored in memory. This comparison imitates the interpretation step of a responsible agent. The outcome of such picture interpretation is based on the

[331] Fujiyoshi, Hirakawa, and Yamashita, "Deep Learning-Based Image Recognition for Autonomous Driving."

‘training’ of the NN in the past by which it has learned a (mathematical) mapping between pictures and objects. Such interpretation of pictures is also far from perfect as the actual picture and its objects (almost) never correspond with the pictures of objects in the database. The latter can be due to many different factors, such as the difference between the actual objects and those in the database, the difference in lightening condition by which the objects in the database are taken, the partial occlusion of the picture, etc. The result is that the outcome is seldom 100% accurate and, in addition, with the outcome a so-called “class probability”, that indicates the probability by which a classified object belongs to a certain class. For example, the classified human in the above imagined picture belongs for 80% to the class of an elderly person, 10% to a motorcycle and 10% to a ‘don’t know’-class.

VII.3.3 Accountability

Accountability is usually defined in a legal context of who can be deemed responsible for possible erroneous actions that cause casualties or damage.

However, in describing the key features of a responsible self, Niebuhr refers to accountability as the third of the four features of responsible actions “insofar they are made in anticipation of answers to our answers.”[332] Therefore, we not only react to interpretations of actions against us, but responsibility means that we are in a dialogue with other agents. In such a dialogue we anticipate a reply, whether this being by “objections, confirmations, and corrections.”[333] For accountability, answers must be given that fit in the whole story. For

[332] Niebuhr, *The Responsible Self: An Essay in Christian Moral Philosophy*, 64.

[333] Niebuhr, 64.

example, in an ADS context, whenever the driver questions the AI agent on what the object is in front of the car, as discussed in Example ***3***, and the AI agent answers that a certain object has been detected, the responsible driver is expected to answer back. This may be either by confirming or even correcting the AI agent. Such accountability would allow to further train the Neural Network using the corrected information.

Next to this anticipation of a reply, being engaged in dialogue also means "to answer questions addressed to us, to defend ourselves against attacks, to reply injunctions, to meet challenges."[334]

The topic of accountability may touch upon other aspects than the action committed by a responsible agent. It may, moreover, address other dimensions of the existence of that agent or even consider a wider social dimension of the society in which the act has been committed. This social dimension brings us to the next point of the responsible self, namely that of social solidarity. This is discussed in the next subsection.

This opening up by looking at other aspects than the action may be helpful, even crucial, in learning from failures. For example in ADS, when a driver is not responding to an alarm of the ADS and thereby causing an accident, one may legally keep the driver accountable. However, in a larger perspective of the existence of that agent, the underlying cause may have been the absence of proper driver training to deal with that and other alarms.

[334] Niebuhr, 56.

VII.3.4 Social Solidarity

The actions of a responsible self are being made in “a continuing discourse or interaction among beings forming a continuing society.”[335] A keyword of importance here is the word ‘continuity’ and this both in the interaction between all parties involved in the action as well as these parties operating in a common society. Discontinuity leads to disconnected interpretations and discontinued actions. Such discontinuity opens up the possibility for a so-called *responsibility gap*.

Recalling the more compact notion of responsibility gap of Sven Nyholm given in Chapter I, but now restricted to a system that is operated by different human agents, a responsibility gap might occur when one of the agents acts outside the control and oversight of another agent operating the same system. The definition Nyholm uses of ‘*Control*’ is the ability to supervise and have full authority over the agent under control. This violation of control and oversight happened, for example, in the accident of the Herald of Free Enterprise before the Belgian coast on March 6, 1987. Then an assistant boatswain, because of drunkenness overslept and thereby failed to close the bow doors on time. The responsibility gap occurred because the captain of the ferry was not informed of the assistant boatswain fallacy but presumed to door was closed as the assistant boatswain had always done before. This responsibility gap caused the death of 193 human lives.

The responsibility gap can occur when automated systems are involved, for example, when it is unclear who of the possible human operators involved is controlling or overseeing that automated system while in the course of actions taken by the ADS casualties result or people are

[335] Niebuhr, 65.

killed. Such a scenario is described in the following example, based on the example given in Nyholm.[336]

***Example 4*:** Consider an ADS that in its automated or autopilot mode can execute the driver's particular plan to go from home to a particular shop. However, the means to achieve the driver's goal is determined by the programmer of the company, who developed the navigation system of the ADS, as the navigation system determines the route to be taken. Now in the middle of the route towards the shop, an accident is caused by the ADS. It is then unclear who is responsible or who should receive retribution for this accident. Did the driver select the destination? Alternatively, is it the programmer who made the software select a particular route on which the accident happened, while alternative routes were possible?

Even when arguing, as done by moral philosophers such as Roos de Jong[337] and Sven Nyholm[338], that ADS cannot be regarded as acting on their own, de Jong shows contrary to Nyholm that responsibility- and retribution-gaps, the latter defined as the desire for retribution without appropriate subjects of retributive blame, remain. Taking Example 3 as a case study, de Jong shows that there are two sets of collaborations involved between humans and ADS. There is the "driver-ADS"-collaboration and there is the "programmer-ADS"-collaboration, each having

[336] Nyholm, "Attributing Agency to Automated Systems," 1211.

[337] Roos de Jong, "The Retribution-Gap and Responsibility-Loci Related to Robots and Automated Technologies: A Reply to Nyholm," *Science and Engineering Ethics* 26, no. 2 (April 1, 2020): 727–35, https://doi.org/10.1007/s11948-019-00120-4.

[338] Nyholm, "Attributing Agency to Automated Systems," 1201.

their goals, and not being part of a one-line command.[339] The lack of such single line command causes a responsibility gap between the two human-ADS collaborations as their independent action may be the cause of the accident, as in the example of the navigation system taking a particular route without the consent of the driver. This shows that imposing a human agent in control over (a part of) an automated system is not sufficient to resolve responsibility gaps. For Nyholm, the responsibility gap could be resolved if there is only a single collaboration between, on the one hand, one human agent or several agents being part of one line command, and on the other hand, an automated machine, when the automated machine collaborates under the supervision and authority of that one human agent.

This could be achieved, using Niebuhr's continuity argument, when there is continuity in the action taken by the navigation programmer and continuity in the interpretation of this action by the driver. This would mean that the driver is being made aware of the choices of the navigation system and at all times can respond by changing the navigation route.

As ADS can execute certain actions, various moral ethicists have tried to investigate what type of agency can be assigned to automated systems. Using a functionalist approach, Nyholm arrives at the fact that current ADS may be assigned with "Domain-specific supervised and deferential principled agency," that is being able to "pursue a goal based on representations in a way that is regulated by certain rules or principles, while being supervised by some authority who can stop the agent or to whom control can be ceded, at least within certain limited domains."[340] The fact that ADS can collaborate gives it a kind of collaborative agency, whereby this deferential and supervised agency is exercised in response to somebody else's initiative. This

[339] de Jong, "The Retribution-Gap and Responsibility-Loci Related to Robots and Automated Technologies," 731.

[340] Nyholm, "Attributing Agency to Automated Systems," 1208.

somebody may be the driver setting the travel goals and the social setting the safety and traffic rules.[341]

Using the elements of Niebuhr's responsible self, such current ADS may be able to imitate the elements of responding in the present to actions of subparts of the ADS (such as sensor readings) or human operators' input, they may do this by interpreting these actions by the intelligence that is stored in their memory (such as in terms of a mathematical model), what is lacking for most present ADS is the ability to act accountably. This is because the current ADS cannot enter in a dialogue on equal footing with responsible humans, to understand (interpret) criticism on its agency, and have the ability to defend or alter its actions based on principled criticism of its agency. For Niebuhr, such principled criticism could be the result of the self's awareness of its uniqueness and its ability to deal with sin, guilt, and salvation. Both require reference to the One Beyond all, as discussed in Section IV.4.5. Also with the insights of Bonhoeffer in Section IV.3.6, the necessity for ADS to be under control as defined by Nyholm on page 151 makes clear that ADS cannot act in freedom. However, as said by Bonhoeffer in that section, obedience without freedom is slavery, the human-ADS collaboration should be seen as a 'master-slave' hierarchy.

The above exposure highlights that the relationship between humans and ADS should be seen as a relationship between humans and technology. In a broader context, we could interpret

[341] Nyholm, 1211.

ADS operating in automated traffic environments[342] as "Active technological environments."[343] Such active technology is becoming a mediating milieu, "merging with the world to the point of becoming invisible, but at the same time intentionally directed at humans and helping to shape how humans act, perceive, and live their lives."[344] Such technology will not take away human responsibility, making them puppets in an orchestrated world, but on the contrary "makes us more responsible."[345] This viewpoint of Verbeek may be interpreted using the insights of Niebuhr and Bonhoeffer. Niebuhr says in characterizing the responsible self that such selves "interpret the things that force themselves upon us as parts of wholes."[346] As technology may seemingly force actions upon its users, following this line of thinking of Niebuhr, it remains important that these users interpret those actions within the whole of their endeavors. So, the use of ADS should not be experienced without being able to interpret its actions. Bonhoeffer makes this even more explicit when he says (see page 66) that responsible humans should discover the intrinsic laws of the things of the world, or designed by the world. Such discovery of the operation of ADS to be able to interpret its actions should indeed make users of ADS more responsible rather than less. When looking up an ADS as a 'thing' I can use, Bonhoeffer's

[342] Xiang Zhang, Wei Liu, and S. Travis Waller, "A Network Traffic Assignment Model for Autonomous Vehicles with Parking Choices," *Computer-Aided Civil and Infrastructure Engineering* 34, no. 12 (July 30, 2019): 1100–1118, https://doi.org/10.1111/mice.12486.

[343] Ciano Aydin, Margoth González Woge, and Peter-Paul Verbeek, "Technological Environmentality: Conceptualizing Technology as a Mediating Milieu," *Philosophy & Technology* 32, no. 2 (June 1, 2019): 322, https://doi.org/10.1007/s13347-018-0309-3.

[344] Aydin, González Woge, and Verbeek, 335.

[345] Peter-Paul Verbeek, "Some Misunderstandings About the Moral Significance of Technology," in *The Moral Status of Technical Artefacts*, ed. by Peter Kroes and Peter-Paul Verbeek, Philosophy of Engineering and Technology (Dordrecht: Springer Netherlands, 2014), 85, https://doi.org/10.1007/978-94-007-7914-3_5.

[346] Niebuhr, *The Responsible Self: An Essay in Christian Moral Philosophy*, 61–62.

warning is in place here to not treat it as an 'idol', that blinds us and endangers our own image of being human.[347]

A final element of the social solidarity of Niebuhr is the Triadic dimension. Next to the responsible self and the other agents the self is dealing with in a particular action, there is always the third reality (see Section IV.4.3). This third reality indicates both the individual reference of the self, as well as the external reference to the self and the other agents represented as "an impartial spectator".[348] In the context of operating ADS, the first reference may be the individual norms and values of the driver, while the second reference is the safety and traffic rules imposed by the society the self and the other agents are operating in. In the context of designing ADS, the first reference may represent the internal compass of values and norms of the designers of such systems, while the second may the code of ethics of the company or technical society these designers are part of.

VII.3.5 Adaptability

Section IV.4.4 summarizes two ways in which Niebuhr provides the responsible self the possibility to break with traditions of the past when these lead to unfitting actions. Recall that in the interpretations the self makes in responding, use is made of information (images) from the past. This in general is used in a defensive way treating people how they have 'always' acted in the past or treating natural forces in a kind of worst-case scenario. This may result in a real cautious behavior that might lead to safe driving, but maybe not efficient.

[347] This topic of idolization of things and the effect it may have in desecrating human beings is discussed in Section IV.3.3.

[348] Niebuhr, *The Responsible Self: An Essay in Christian Moral Philosophy*, 84.

The first way of adapting your response is when forces of nature, your interpretation of these forces, is giving rise to unfitting actions to the reality you are acting in. To change this old pattern of response, radical doubt needs to be expressed. Therefore, to be able to make changes in the interaction between agents, these agents must exchange information about the confidence they have in a particular action. In example ***3***, the class probability by which a NN expresses its confidence in an object classification outcome could be used as such an expression of confidence. Based on this confidence level the human driver could respond responsibly when the class probability drops below a defined threshold, defined by the manufacturer. Taking the continuity aspect as highlighted in VII.3.4 of the topic of Social Solidarity into account, it is necessary to continuously update these class probabilities and issue an alarm when it drops below the threshold. When a human operator is trained to understand the meaning (seriousness) of such indication it will enable him or her to respond. This example demonstrates that by providing the human operator regularly with understandable information to appropriate the 'what is going on?' question, an improved user responsibility will result.

The second way of adaptation that Niebuhr suggested was in the response to persons or communities. Such adaptation should occur less frequently compared to the first way of adapting. This second way is of interest in the adaption of the interaction between the developer(s)/programmer(s) of the ADS and the human driver. Now such interaction is non-existent, as the human driver treats ADS as a black box. Following the simple defense ethic Niebuhr highlighted in Section IV.4.4, the simplified response can then either by one of having full confidence in the ADS, a kind of blind confidence, as was the case for the Tesla-S driver in Example 2 with fatal outcome, or having no confidence and thereby simply not using ADS. A better way to adapt the (response-action) interaction between ADS developers and the human driver is to create the possibility for the ADS developers to tap into the expertise of the driver. Such expertise is based on the driver's empirical conscience in his/her activity of evaluating the fittingness of an action. In this evaluation, their own standards can be used, as well as the criticism or praise they have received from society, for example, during training. Based on this expertise, and restricting to the ADS object recognition task, the discrepancy between the ADS and human driver needs to be logged whenever the ADS has resulted in a class probability below a threshold. This requires a dialogue between the developers of the ADS and the human driver. The specific form of such a dialogue is not further elaborated on and is left to future research. The important thing is that such a dialogue may result in the improvement of the object recognition by training the AI with the 'new' data that was badly classified. Such interaction should be invoked each time a bad classification is made, and not only when a bad classification has resulted in a fatal accident, as was the case in the Tesla-S fatal accident reported in Example 2.

Whenever the human driver recognizes that his/her reporting of such bad classification leads to improvements of the ADS, re-interpretation of the relation between ADS developers and human drivers will result. In such reinterpretation one may need to deal with the criticism, the need to defend or alter one's action in dealing with the reactions, in the interpretation of the answers to our answers of companies, and in dealing with the third reality as discussed in Section IV.4.3. This indeed may result in an increase of doubt to both the ADS developer and human driver in the short run, but when updating our older images results in better fitting responses in the future, human confidence (in handling that particular action) is increased. The outcome will in general by a safer ADS.

VII.3.6 Dealing with Personal Faults

In general, the personal dimension of traffic accidents is dealt with in a legal framework using tort, liability, or strict liability legislation for traffic law (a division of criminal law), as well as for civil law, in case of handling construction failures. There are two challenging problems with a clear ethical dimension in relation to handling failures. The first is that the outcome in a particular accident may not always be conformal to common-sense morality. The second is the guilt related to an accident even after the accident has been properly dealt with legally.

An example of the first is the ADS accident with the Tesla-S model, explained briefly in Example 2. Here the fact that the driver was informed in his contract when buying this Tesla-S model to always remain vigilant and ready to regain manual control over the vehicle conflicted with the selling argument of a car with ‘autopilot. The consequence was that he was too late to recognize the collision with the truck in front of the car, and the AI system of the ADS misclassified it as a white cloud. Tesla admitted that their product made a failure as they promised to update the software after the crash. They “admitted to [having] been causally responsible, at least in part, for the fatality, though they denied that they were to be assigned legal, and perhaps also moral, responsibility for what happened.”[349] This flaw is a consequence of handling the accident via what Niebuhr called the principles of ‘man-the-citizen’. Moreover, the discrepancy between the legal outcome and the causal result is due to neglecting the actual event and its responses or the lack thereof. While Tesla promised to update its software, this case highlights the distance that deontological ethics creates from actual events by hiding behind principles and rules.

[349] Nyholm, “Attributing Agency to Automated Systems,” 1202.

The second challenge of guilt is not addressed in a legal framework and is left to the individual. In this area, the ethics of responsibility of Bonhoeffer and Niebuhr can make a difference. Continuing the Tesla-S example of the previous paragraph, according to Bonhoeffer's *Willingness to Actively Embrace Guilt* of Section IV.3.5, Tesla did not demonstrate an act of responsibility by seeking self-righteousness through justification by following a principle. The search for self-righteousness via principles might have been successful for the Tesla manufacturer (legally); it may cause guilt among human agents who have participated in the development of the ADS for this Tesla model. Both the developers of the ADS as well as the management pushing the technicians to meet stringent deadlines to be first on the market might suffer from guilt because of the effect of their involvement. For dealing with such guilt, living in full dependence on God's grace is the solution according to Bonhoeffer, as explained in Section IV.3.4. Redemption offers liberty and the process of healing and renewal can start, as explained based on Niebuhr's responsible (see Section IV.4.5.2). A challenging question could, therefore, be, "Whether a redeemed designer who has been involved in malfunctioning of AI software, is able to function better as a responsible human after having experienced God's grace?"

VII.4 Concluding remarks

The framework outlined in this chapter has provided a critical analysis of the MHC approach of Santonio and van den Hoven using Niebuhr's man-the-maker philosophy. Furthermore, by using the insights of the Responsible Self, a new alternative model is presented that improves human action relating to ADS and makes it act more responsibly. A sketch of a concretization of this new model is presented in the next chapter.

Chapter VIII

Towards Concretizing Responsible Human Control over ADS

VIII.1 Introduction

The six elements of the Responsible Human Control over ADS outlined in the previous chapter are used as a basis to concretize this framework to enhance the responsibility of the human driver while preserving other important aspects such as safety, mobility, and performance. The key to concretizing the framework is in information exchange between ADS and the human driver. In this chapter, I describe the information that can be exchanged to enhance the responsibility of human beings, how that information can be exchanged, and why it should improve the operation of the human operator and the ADS.

VIII.2 Designing for Responsible Human Control over ADS

VIII.2.1 Proposal for a Concretization

In this section, I make several of the elements of Niebuhr's and Bonhoeffer's ethics of responsibility more concrete in an attempt to increase and improve the responsibility of the human driver over ADS. I restrict myself to dual-mode driving systems at the SAE level 3 ADS. Such systems provide partial driving autonomy in certain circumstances, such as when driving on highways. During such "auto-pilot" mode, as it was indicated by the Tesla manufacturer for its Tesla-S model, the driver needs to constantly remain vigilant and ready to regain operational (manual) control at any time.

The concretization of several elements discussed in the previous chapter are presented using Michon's three functional levels of human behavior in controlling cars.[350] The three functional levels are the strategic, tactical, and operational levels, which are defined in Section II.2. They are depicted by the three boxes in Figure 2. When applied to a car without ADS, both blue and black boxes of the same functional level are merged and executed by the human driver.

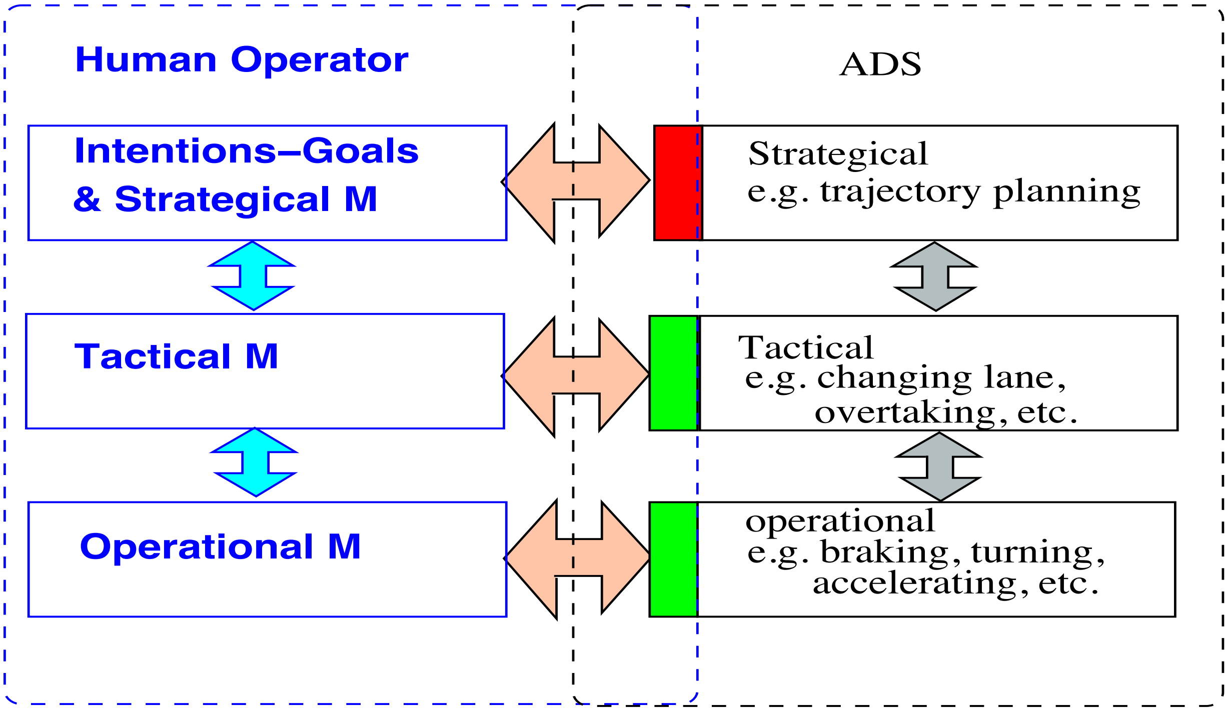
Human Operator
Intentions–Goals & Strategical M
Tactical M
Operational M
ADS
Strategical e.g. trajectory planning
Tactical e.g. changing lane, overtaking, etc.
operational e.g. braking, turning, accelerating, etc.

Figure 2: [Black dashed box] Different functional levels of Control of a driver according to Michon, possibly taken over by the ADS as indicated by the red or green color, with green indicating that ADS is active and red that it is not. [Blue dashed box] The Human operator activity.

With the presence of ADS, the latter system could in principle, depending on the level of ADS as presented in Chapter II, take over all three functional driving levels of Michon's model. I

[350] Michon, "A Critical View of Driver Behavior Models," 485.

duplicated these three functional levels for the ADS in Figure 2. In this figure, the black part represents the automated part, which is indicated by the ADS part, and a blue part represents the human driver's activities. This figure indicates that automation is possible on all functional levels, as would be the case for SAE level-3 for ADS. The grey and blue arrows between these boxes represent the exchange of information between these functional driving levels. Most of this information is implicit and hidden. For example, the data exchange in ADS between tactical and operational levels goes via an internal car communication network, such as the "CANbus".[351] The exchange of information as indicated by the pink arrows is between the human driver and ADS, and it will become crucial in enhancing the responsibility of the human driver over ADS. However, not all levels need to be active all the time in parallel. It might be as illustrated in Figure 2, where only two levels, indicated by the green indicator, are active, "while leaving the driver with strategical control."[352] This particularity is omitted from the discussion and I simply consider a dual model ADS system as defined in II.3.

The accident with the Tesla-S model[353] illustrates the failure of the deontological approach, as clarified by Niebuhr's man-the-citizen symbol. While Tesla warns the driver that he or she is always responsible for the safe operation of the vehicle,[354] clearly trying to impose a rule on the driver, it bypasses the self-defining nature that can only be experienced when man is dealing with his limits in real-life events. The attempt to resolve this man-the-citizen ethics

[351] Premio Inc, "CANBus: The Central Networking System of Vehicles," Premio Inc, accessed 20 June 2019, https://premioinc.com/blogs/blog/can-bus-the-central-networking-system-of-vehicles.

[352] Mecacci and Santoni de Sio, "Meaningful Human Control as Reason-Responsiveness: The Case of Dual-Mode Vehicles," 105.

[353] David Shepardson, "Tesla, NTSB Clash over Autopilot Investigation," *Reuters*, 12 April 2018, https://www.reuters.com/article/us-tesla-crash-autopilot-idUSKBN1HJ2JS.

[354] Shepardson.

approach with the man-the-maker alternative demonstrated that this might result in enhanced safety but at the cost of (dramatic) performance breakdown (see Section VII.2). Using man-the-answerer's approach, a crucial element to invoke man's responsibility is the exchange of information, making this information interpretable and invoking man's ability to respond. The exchange of such information is indicated by the pink arrows in Figure 2.

By interpreting Niebuhr's elements of response as action operational, key questions need to be asked and answered. The initial questions, which aim to build confidence between the ADS developers/programmers and the human driver, are about "What?" and "What is going on?" However, it will be shown that that confidence level can further be improved by additional how-, and why questions. In the concretization of these three questions and their answers to actively involve the human driver in ADS, I start with the what-questions.

A major problem with many current AI solutions is that they act as black-boxes, with the user having no information about what is going on. This was the case with the reported Tesla-S accident, where the human driver did not know about the NN classifier difficulty of incorrectly classifying a truck as a cloud. Instead, a generic warning was regularly issued. It is well known that when operators regularly experience a particular alarm without consequences, the so-called "alarm-fatigue" occurs.[355] The ignoring of alarms is further worsened by the impression created by commercials promoting the Tesla-S model, with the driver "taking a nap as his car navigated busy traffic."[356]

[355] Naveed Saleh, "The Dangers of Alarm Fatigue," *Psychology Today*, January 12, 2018, https://www.psychologytoday.com/blog/the-red-light-district/201801/the-dangers-alarm-fatigue.

[356] Yardon and Tynan, "Tesla Driver Dies in First Fatal Crash While Using Autopilot Mode."

VIII.2.2 Addressing the “What-is-going-on” questions

For the sake of brevity, I restrict the discussion of this topic to the object recognition task of ADS. Generally, object recognition tasks from an image consist of several tasks. These tasks were listed in the Introduction on page 14. These tasks already clarify that the objective of deep learning-based image recognition addresses many interesting “What?”-questions of primary interest to man-the-answerer ethic. The main problem is that this information is hidden from the driver. This causes a disconnection between the human driver and the ADS and in case of an accident like that with the Tesla-S also a responsibility gap.

The call for an explanation of what AI is doing in the case of object recognition is recently addressed via the introduction of so-called attention maps. An example of this new development is depicted in Figure 3. Here the caption is based on comments to Figure 12 of

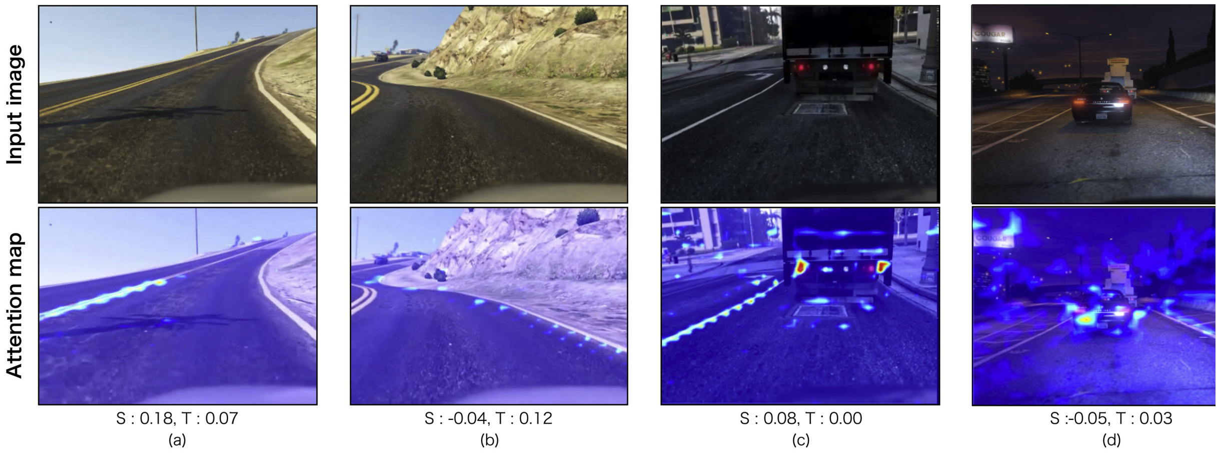

Figure 3 [Figure 12 of Fujiyoshi et al.]: Visual explanation in terms of a so-called attention map (lower row) as output of the deep learning of the NN. This attention map is presented in relation to the true scene in top row. (a) the scene in here shows a road curving to the right with a strong response to the center line of the road as indicated by the illumination of that center line. The steering value on the bottom of the attention map, indicated by the "S" value of 0.18 indicates through its positive value a steering to the right. (b) this is somehow the opposite of (a) with a steering to the left and curving to the left, (c) the attention map draws attention to the stop lights of the truck in front and the zero value of the throttle "T" indicates that the ADS is closely watching the vehicle ahead and (d) indicates a strong response of the ADS to the car ahead because lack of knowledge what happens in front of that car.

Fujiyoshi et al.[357] Presenting such information to the driver, even when he is not requested to take control of the vehicle, provides useful information to allow that driver to interpret what the ADS is doing. It enables the driver to develop an understanding of the object classification by comparing the ADS interpretation with his or her own interpretations. As discussed in Section VII.3.5, such comparison between 'images of the past' and 'new ones' is a resource for adaptation. In such situations, errors can be made as well, as improvements are possible when learning from these errors. The latter aspect is picked up in the analysis of the 'how question' in the next subsection. However, the comparison on the one hand creates confidence between the

[357] Fujiyoshi, Hirakawa, and Yamashita, "Deep Learning-Based Image Recognition for Autonomous Driving," 251.

driver and the ADS when it results in frequent agreements. On the other hand, it leads to concern and automatic alert when discrepancies are high. Therefore, it may be concluded that this extra information invites in a natural way to stay alert on the driving action, though the ADS might be doing most of the work. Therefore, when in case the discrepancy in the comparison is too high it will motivate the driver to interfere. Of course, the driver must learn to make use of the information supplied by such attention maps and to respond to it adequately. However, it demonstrates that the new technology of ADS does not call the driver to be passive. The fact that additional information is being presented to the driver, under the motto that two pairs of eyes are better than one, will contribute to increased safety of driving, but not necessarily reduce the driver's workload. The information between the ADS and the driver is bi-directional as indicated by the pink arrows in Figure 3. Though most of the info is from the ADS to the driver in presenting and updating the attention maps, the fact that the driver is on the alert all the time may cause him or her to interfere and take over the driving activities in case of perceived too large discrepancies between the presented information and his or her assessment of the situation.

VIII.2.3 Addressing the "How?" Questions

The how-question from the human driver perspective is not how the machine learning methods work, but how its results need to be interpreted to prepare for a response, an answer, a further investigation. In addition, this element is missing in the current ADS. A possible way to make such information available in object image recognition is via the use of the class probability. This metric was introduced in Example ***3*** to indicate the confidence level in the outcome of a classified object from images. Such information can be readily added to the attention maps in Figure 3, for example, by adding the class probability to the (most) relevant objects, in other words, those objects that are illuminated. Using Figure 3(c) when a cloud is detected instead of the truck, the class probability of the truck wrongly classified as cloud could start flashing (in a red color). The driver could then interpret this information to either respond by taking over the driving at the operational and tactical level or by doing nothing when a false alarm is observed. These skills of mastering the interpretation and response need to be taught in an ADS-training school to be discussed in Section VIII.3,

VIII.2.4 Addressing the "Why?" Questions

These questions are often the most difficult ones. This is because it touches upon the intrinsic values and norms a user or a developer has concerning the thing or situation at hand. Nevertheless, such questions are highly relevant for the adaptation of the responsible self and the ADS. They facilitate learning from errors made and allow innovation. To illustrate this in the context of ADS, let us again consider the misclassification in the Tesla-S example.

When, as in Section VIII.2.3, the driver realized that the detected object that was illuminated in Figure 3(c) was not a cloud but a truck, and after having taking over the driving from the ADS and safely continuing the ride, the 'why' question can be addressed. In general, such a question should be addressed by the designer of the object image recognition part and stimulate him, her, or the team to look for answers to improve. In this search, the new data of the driver will be used, at least when he or she has classified the object correctly, to improve the machine learning database. Here one could ask the question: "Why is the machine learning methodology not able to do this on itself?" The reason for this can be explained using Niebuhr's model of the responsible self. Similarly, to the use of 'old images' as a reference framework, as indicated by the third reality in Section IV.4.3, machine learning algorithms are 'stuck' by these old images when being confronted by new information. That is, when the new event is not present in the memory of the (current) machine learning algorithms, it will never be able to relate to this. When the class probably is too low, the best option might be to label it a 'don't know class', stop the vehicle safely, and start finding information to learn what it was. Learning requires making errors, being stuck, and using your innovation skills, such as highlighted in Section VII.3.5, to update the old memory.

VIII.3 Responding and Adaptation within ADS

The previous section highlighted that the application of the elements of the man-the-answerer ethic to ADS increases the intensity of interaction between the human and the ADS, instead of decreasing it. The latter is generally used as a selling argument, like taking a nap while driving in busy traffic, as mentioned at the end of Section VIII.2.1. However, in line with Peter-Paul Verbeek's observation that when "seeing the moral significance of technologies makes us more responsible, rather than less,"[358] using ADS responsibly should indeed increase the activity of the driver in general, as well as driver alertness at the moments it matters most for maintaining safety. This increased action-response activity of the human driver, according to the man-the-answerer model of Niebuhr, will increase human responsibility for ADS, while ADS may do most of the driving work. The overall benefit could very well be increased traffic safety.

However, the responsibility of the human controller over ADS comes at a cost: new skills need to be learned and acquired. Therefore, I present several elements that allow both the human operator as well as the human developers to adapt their old response patterns to learn new ones that *better fits reality*.

While the information provided in the previous section may be of great value for acting responsibly, the key requirement in Niebuhr's model is still *to respond* to that information. The

[358] Verbeek, "Some Misunderstandings about the Moral Significance of Technology," 85.

additional information that the human driver will receive in the new proposal made in Section VIII.2 is new and maybe dense. Therefore, responding to this information requires the development of a sound understanding of the function, capabilities, and limitations of the ADS. Furthermore, to take ownership of the actions requires building up the skills to assess appropriately and efficiently the provided new information and knowing what the best fitting action is to respond to the assessed information. The last two requirements require additional training of the human driver. Moreover, when learning to deal with the why's as outlined in Section VIII.2.4, learning to properly exchange information between drivers and the designers of the ADS is of great interest to improve the system, in other words, to ensure the ADS 'learns'.

Learning to operate automated systems requires dedicated attention and time. In this sense, commercials that promote ADS as relieving the human driver from most of the driving responsibility are misleading. This is in line with the common assumption that "automating jobs makes them easier."[359] However, "experience in the field of aviation shows that automation sometimes complicates things in unexpected ways."[360] Training people for operating automated systems is a field of study that has created a significant body of research results that can be used to learn to use automation effectively and safely. This experience should be used to learn/train the human driver to exploit the capabilities of ADS but also to learn to interpret and respond to its limitations. Therefore, the learning strategy forms an integral part of the new Responsible Human Control over ADS strategy presented in this thesis.

[359] John Barnett, "Training People to Use Automation: Strategies and Methods," *Journal of Systemics, Cybernetics and Informatics* 3, no. 5 (October 1, 2005): 73.

[360] Barnett, 73.

How such learning and information exchange between ADS, human drivers, and developers should be done, and how issues surrounding the privacy of that information should be dealt with, are beyond the scope of this thesis.

The next section addresses the possible resistance or opposition that modern Western people may have to this increased workload.

VIII.4 The blind-spot in the Western modern mind resisting the new approach

In the discussion of Niebuhr's view on the human consciousness in Section IV.4.7, I stated Paul Lehmann's conclusion that Kant's work on the conscience opens the door for subjectivity and individualism. In this section, I evaluate how both traits hamper people's acceptance of behaving more responsibly when using active-high technology, such as ADS.

The individualism of the modern Western self is generally encouraged by active high technology on three different levels: the desire for social recognition, authenticity, and the rise of instrumental reason. The discussion on how these three levels prevent acceptance of the new proposal presented in this chapter is based on Carl Elliott's work on "Enhancement Technologies and the Modern Self."[361]

Modern Westerners like to be recognized and respected for who they are as individuals. This, however, has not always been the case. In the 18th century, public recognition of a person's individual identity was based on the person's social origin. While social roles still influence a person's status, such recognition is not merely given and can partly be earned or generated by the

[361] Carl Elliott, "Enhancement Technologies and the Modern Self," *Journal of Medicine and Philosophy* 36, no. 4 (August 2011): 364–74, https://doi.org/10.1093/jmp/jhr031.

individual. That earning, however, is not something the individual can decide upon privately, as Charles Taylor writes:

> My discovering my identity doesn't mean that I work it out in isolation, but that I negotiate it through dialogue, partly overt, partly internalized, with others. This is why the development of an ideal of inwardly generated identity gives new and crucial importance to recognition. My own identity crucially depends on my dialogical relations with others.[362]

Though the link with others is mentioned, this relational movement is clearly ego-centric to receive recognition about the "I", the other acting as a 'servant' for the growing self-esteem of the "I". Such striving might be through gaining honor through extraordinary personal achievements, such as winning a gold Olympic medal; in our social structures of social equality, honor has been replaced by dignity. In my opinion, Elliott correctly remarks that "dignity is enabled to everyone, but not everyone has it."[363] This lacuna that humans experience in our modern world is used, or misused, by the commercial industry to seduce people into acquiring that 'one thing' that will make them a valuable self. In this way, we may interpret the mentioned slogan that promotes self-driving cars as vehicles that allow you to take a nap while maneuvering through busy traffic as an appeal to that feeling of exclusivity that will give you the feeling of being someone. Thus, ADS technology becomes more of a status symbol than a technology that aims to improve safety, protect the environment, etc. The consequence of such seductive slogans is that humans will not be eager, or even willing, to act responsibly as outlined in the new proposal. This insight could also follow from Bonhoeffer's warning about 'worshipping' things as false idols, as discussed on page 63.

[362] Charles Taylor, *The Malaise of Modernity* (Concord, Canada: Anansi Press, 1991), 47–48.

[363] Elliott, "Enhancement Technologies and the Modern Self," 367.

The second element of individualism is the authenticity of the self. Modern men often use the language “to get in touch with the true self.”[364] Taylor describes the ideal of authenticity as follows:

> There is a certain way of being human that is my way. I am called upon to live my life in this way, and not in imitation of anyone else’s life. But this notion gives new importance to being true to myself. If I am not, I miss the point of my life; I miss what being human is for me.[365]

This view of authenticity is very much in line with Kant’s interpretation of the human conscience, as discussed in Section IV.4.7. The authenticity as stated by Taylor again boosts the thirst for ego-centricity, where it is not the self-caring for the other, but it is the “I” caring about his or her own life. Elliott emphasizes that this self-determining freedom moves a level beyond what is sometimes called negative liberty or liberty to do what you want without external interference. This sealing of the individual from the external world is wonderfully portrayed by slogans about self-driving cars that you can take a nap while driving. It creates the ultimate desire for the authentic self to be freed of external pain caused by other traffic participants on the way you drive your self-driving car, and have time ‘to be your true self’. When being confronted with the limits of the technology, such desires will turn into an illusion. However, then it is often too late as illustrated by Examples 1 and 2, with dramatic consequences.

The third element of individualism is *instrumental* reason. In our modern society, technology has created the impression that the world is something to be molded and shaped by the ‘free’ will of humans. Technology has, to say it with the words of Max Weber, disenchanted the world. The combination of scientific methods and enlightened reasoning has made the world

[364] Elliott, 369.

[365] Taylor, *The Malaise of Modernity*, 29.

into an instrument of human design instead of an object of mystery and reverence.[366] Charles Taylor uses to phrase "instrumental reason" to indicate the kind of rationality humans draw on when they calculate the most economical application of means to a given end.[367] Such instrumental reason values the world by measuring it in terms of what it can provide, and looks at the world to be redesigned for the well-being and happiness of the individual rather than viewing the world as a sacred structure, grounded in the order of things or the will of God.[368] Instrumental reason creates the illusion that humans can be sealed from the unhidden in the face of chance and accident. It provides a kind of conviction of infallibility that boosts human pride. However, it creates a kind of blindness and lack of readiness in case the world resists being completely controlled. Such blindness to the imperfections of the ADS technology contributed to both accidents with ADS reported in Examples 1 and 2. This illustrates that this third element of instrumental reason also contributes to the individualistic, ego-centric attitude of possible users of ADS. Rather, it may contribute, as a generalization is out of the question, as there may still be users who, with good intentions, will demonstrate an attitude of humbleness and altruism to care about the other.

VIII.5 Engaging in dialogue with Bonhoeffer and Niebuhr once more

How can modern man be helped to overcome the blind spots that were highlighted in the previous section? Part of the answer lies in balancing their individualism with collectivism so that they again start living in Niebuhr's triadic structured relationship. The three elements of

[366] Elliott, "Enhancement Technologies and the Modern Self," 371.

[367] Elliott, 371.

[368] Taylor, *The Malaise of Modernity*, 5.

Bonhoeffer's theological anthropology, as summarized in Section IV.3.8 on page 71 concerning his doctoral dissertation *Sanctorum Communio* (SC), can deliver modern man from the captivity of individualism.

Aligning humans and their created artifacts with Jesus reveals that nothing that is designed and operated by humans, also AI for ADS, will be perfect. This realization invites humans to discard pride and feelings of infallibility. Dealing with failures might give rise to guilt, as discussed in Sections IV.3.5 and IV.3.9. However, every human can rely on restoration through Jesus Christ. The second element was Bonhoeffer's appeal to personalism. The relational element of the I-Thou encounter invites users of ADS to not consider the other they are encountering on their traffic itinerary as another ADS, or as an intruder to their comfort zone, but as another person created by God. Here, Bonhoeffer's vicarious representation urges users of ADS technology to stand in the footsteps of the surrounding traffic users and the other passengers in the car. The third element of Bonhoeffer's theological anthropology, namely the collective-personhood restored in Christ, might be the most difficult to extrapolate to the modern mind. However, in this case, the invitation to think about the collective would place a special focus on the most vulnerable ones in traffic scenarios, such as the elderly man or woman crossing the crosswalk slowly, or children playing on the sidewalk, or perhaps the one inattentive woman that jaywalks when you are in a hurry.

Therefore, the increased freedom gained by using active high technology such as ADS does not allow you to relinquish your responsibility as a human driver, but it should be used to act more responsibly and care for the other.

Chapter IX

Conclusions

In the Introduction, I formulated the following research question: *How can the theological ethics of responsibility, especially as developed by Dietrich Bonhoeffer and Richard Niebuhr, be integrated with existing teleological and deontological frameworks of ethics, for the evaluation of the use and development?* I further defined this research question in two sub-questions to make the development and usage aspects more precise. The first sub-question reads: *How could we integrate ethical decision-making when handling life-threatening traffic dilemmas as well as in mundane traffic operation?* The second reads: *When considering human-AI interaction, how should one enhance, or make possible, that humans can take responsibility even when the ADS does most of the work?*

In the first two chapters, I clarified two important notions in this research question. In Chapter I, the notion of ethics of responsibility related to AI was explained as "Ethics of Responsibility for AI is ethical research or investigation that systematically addresses moral questions under the guiding principle of responsibility in the context of AI." Here, use was made of an extensive study of the EU task force of the Centre for European Policy Studies (CEPS) and their report on "Artificial Intelligence: Ethics, Governance, and Policy Challenges."[369] This report, formulated from a deontological ethical perspective, offers guidelines for the designer and user of AI to make AI trustworthy or to enhance its trustworthiness. These guidelines help formulate a code of ethics for companies and organizations, but they are too generic and vague to address concrete ethical dilemmas. The CEPS report of the EU task force suggests first

[369] Renda, "Artificial Intelligence: Ethics, Governance, and Policy Challenges."

conducting a concrete additional analysis for concrete ethical challenges, and second, finding complementarity between the AI system and the human user of it. Thus, the report does not aim to exclude the human user of AI or that the human user should use AI blindly. Following these two recommendations of the CEPS report, I focused on the design and use of AI for ADS and selected a level of automation, namely level-3 ADS in the taxonomy of the Society of Automotive Engineers (SAE), where the driver of the AI technology has to continue to act as a fall-back of ADS during its operation.

The first sub-question is analyzed in Chapters III and V. In Chapter III, I present an overview of the way ethics of AI for ADS was classically handled mainly by philosophers. In this classical way, hypothetical traffic dilemmas were formulated that called for a selection about which life to save by applying normative ethical frameworks. These frameworks were either teleological or deontological and aimed for what I called *programming ethics* in ADS.

To bring theology in dialogue with the mainly secular approaches that address the ethics of responsibility for ADS, I reviewed the works of Bonhoeffer and Niebuhr in Chapter IV. In this chapter, the alternative of both theologians for teleology and deontology is formulated. For Bonhoeffer, this was his structure of responsibility in his major work *Ethics*,[370] while for Niebuhr it was the man-the-answerer synecdochic analogy described in *The Responsible Self*.[371] The dialogue resulted in four different contributions.

First, the ethical frameworks of Bonhoeffer and Niebuhr were used to criticize the classical ethical approach for programming ethics in AI for ADS. This classical approach was the preferred approach by philosophers addressing the ethics of AI for ADS. The critique

[370] Bonhoeffer, *Ethics*.

[371] Niebuhr, *The Responsible Self: An Essay in Christian Moral Philosophy*.

consisted of two major parts: (a) The lack of realism of these highly hypothetical traffic dilemmas. Not only is the probability that they will occur in reality very low, but the lack of realism also stemmed from the fact that all, albeit few, of the considered facts in the traffic dilemma were assumed to be exactly known. In reality, AI has to deal with uncertain data, such as when its object detection software is not able to discriminate between a white cloud or a white van for the object in front of the car, as was the case in Example 2. (b) The neglect of the other as a responsible other in the derivation of the ethical rule. In reality, these others respond to actions and this action-response needs to be taken into consideration.

The classical approach that attempts to program ethics in AI for ADS cannot be considered as a systematic approach for this task. This is because it is not able to come up with a unique rule for one particular traffic dilemma as the outcome depends on the individual making the rule as well as on the normative ethical framework being used. To overcome the two major critiques of the classical approach and its inadequacy for programming ethics, Chapter V reviewed recent engineering developments that emphasized enabling human designers of AI to *program* ethically instead of programming the ethics in AI.

The *second* contribution of this thesis was to understand that these engineering contributions brought this approach of programming ethically more in line with Bonhoeffer's structure of responsibility and especially with the vertex of acting in accordance with reality (*Wirklichkeitsgemäßheit*). As recommended by that vertex, ethical programming enables the programmer to consider the actual traffic scenario when determining the action of the car. It enables the determination of this action by previewing a short period in the future of the effect of that action before taking the action, which allows it to take the uncertainty of the facts used to determine the action into account, and the other traffic participants relevant to the actions are also regarded as responsible agents. The approach allows one to integrate teleological and deontological ethics into the design criterion or utility function that is optimized to determine the action. This real-time optimization for the action taken at every decision moment, which might be at the millisecond level, ensures that the 'best' action is taken for the current traffic scenario. Life-threatening and mundane traffic scenarios are considered. The real-time optimization of the utility function that integrates both teleological and deontological ethics might make it possible to violate deontological rules, for example, when crossing a white line in case of an emergency to give priority to a rescue vehicle to free the part of the road needed by that vehicle. The integrative approach calls for the active involvement of both the design engineer and ethical expert in fine-tuning the utility function in the iterative process to evaluate the overall design. While the engineering approach mainly addresses one of the vertices of Bonhoeffer's full structure of responsibility, the analysis has shown that this recent engineering approach paves the way for engineers and theological scholars in ethics to collaborate from the onset of the design of AI systems. Therefore, this second contribution must be seen as an encouragement to theologians and engineers, both with expertise in ethics, to work together instead of working from behind the walls of their communities where they might feel comfortable. Thus, this second contribution calls for true interdisciplinary collaboration.

From Chapter VII onward, I addressed the second sub-question of this thesis. A challenging aspect here is the so-called responsibility gap between the ADS and the driver or human operator involved. This gap occurs in the situation where, for example, due to the uncertainty in the data that the AI system receives, that system remains undecided while, under these circumstances, the driver does not interfere based on his or her perception about ADS. For example, inspired by the commercial that promotes this technology, the driver can take a nap while the ADS is doing the driving. In an attempt to overcome these responsibility gaps, the work on concretizing the concept of Meaningful Human Control (MHC) by the group of Prof. van den. Hoven of the Delft University of Technology is reviewed in Chapter VI. This approach aims to guarantee at each moment that ADS is active the following two conditions: (a) The AI system is responsive to relevant human reasons to act or to refrain from action. This is called the so-called tracking condition. Furthermore, (b) the identification of one or more human agents within the design of the system and its operation who can understand the (remaining) capabilities of the system while being able to appreciate their moral responsibility for the system's behavior. This second condition was called the tracing condition.

In the *third* contribution of this thesis, *The Responsible Self* of Niebuhr was mainly used to highlight the deficiencies of the concretized solution to MHC to ADS. This is done in the introductory section of Chapter VII. The major discrepancy between the work of van den Hoven's team and that of Bonhoeffer and Niebuhr was the way human beings were viewed. The view used in the derivation of the concretization of MHC was a reductionist view, which assumed that humans determine their action by an internal mechanism of evaluating reasons. The concretization then attempts to imitate such mechanisms. Contrary to this reductionist view of human beings, Bonhoeffer and Niebuhr take a holistic view of human nature and its relationship to its Creator and His Son Jesus Christ. This opposite view on the human nature 'naturally' leads to different solutions for the ethics of responsibility in AI for ADS. While the concretization of MHC makes the human a "slave" in its design, who should be goaded to be alert all the time, the holistic view makes the human a "master" of the design and operation, anticipating that the system might have shortcomings or even demonstrate failures.

These opposite views resulted in the *fourth* contribution of this thesis. The application of Niebuhr's model of the man-the-answerer *The Responsible Self* was used as a blueprint to propose a new alternative to make the human driver or operator responsive and, hence, responsible even when ADS is doing most of the work. This blueprint is outlined in Chapter VII. The first attempt to concretize this blueprint is presented in Chapter VIII. Niebuhr summarizes his model as follows:

> The idea or pattern of responsibility, then, may summarily and abstractly be defined as the idea of an agent's action as a response to an action upon him in accordance with his interpretation of the latter action and with his expectation of response to his response; and all of this is in a continuing community of agents.[372]

The four elements in this model are (1) response in the present, (2) interpretation to determine the fitting action, (3) accountability in the dialogue with others, and (4) social solidarity. These four elements are translated into the new blueprint. The first element is translated in the continuous stand-by of the human operator ready to respond to an action, either of the ADS or other agents in the vicinity. The second element requires that the human operator can get answers to questions such as "What is going on?" and "What is being done to me?" before the question "What shall I do?" To realize this aspect of interpretation, it has been shown that the AI system in the ADS provides information on the confidence in its decision making, for example, by presenting the confidence levels by which it has classified objects in front of the vehicle. For the third element, a dialogue must be conducted between the human operator and ADS with mutual anticipated replies. This anticipation of a reply creates an important feedback loop between the human operator and ADS, which makes the system robust to anomalies. When, for example, ADS is expressing its inability to decide due to low confidence in its object

[372] Niebuhr, *The Responsible Self: An Essay in Christian Moral Philosophy*, 65.

recognition, a reply by an accountable human operator may be the overruling of the ADS and taking over the vehicle control. In this way, accountability improves the system's resilience to errors. Finally, the element of social solidarity is invoked by assuring continuity in the interaction between all parties involved in the action. This element reduces the responsibility gap. The combination of these four elements helps to assess he whole system, vehicle-ADS-and-human-operator, to determine the best fitting action to deal with the actual driving scenario. That best fitting action may either occur through the optimization of a utility function as in the approach to program ethically, which was reviewed in Chapter V, or by the human operator based on his experience. Adding to these four elements the elements of adaptability and the possibility of dealing with human errors, I created a system that will be able to learn from its failures while avoiding catastrophes. On the one hand, this learning capability will make continuous improvement to the AI technology possible and, on the other hand, it requires training of the human operator to be able to provide this missing link in the learning chain. However, such training is but a normality, as we can see from airline pilots who have to undergo intense training phases regularly even though they fly highly automated planes. In the concretization of the above-outlined blueprint, as done in Chapter VIII, it has been shown in several ways how one can realize this blueprint by using existing or recently developed technological solutions.

The above four contributions show how the dialogue between theological ethics of responsibility as developed by Bonhoeffer and Niebuhr and secular approaches that are mainly based on teleological- and deontological ethics, contribute to the evaluation of embodied AI for ADS. It has been shown that theologians with expertise in ethics can contribute to an interdisciplinary team of developers of AI from the onset of the design of these systems.

Moreover, this can take place in two ways: First, by being an ideal sparring partner to help fine-tune engineering developments that aim to program AI ethically and, second, by devising and deriving new frameworks that improve the responsiveness and, therefore, also the responsibility of the human operator involved.

These contributions have shown that both in the development of AI as well as in its use, the role of the human operator should not be decimated, but on the contrary, should be stimulated and enabled to act responsibly. While the system around the human operator may provide the means to make this possible, it is still up to the human in his freedom to accept this responsibility. As indicated in Section VIII.4, ‘natural’ barriers might arise in the mind of the modern Western self due to the much-cherished individualism and the strong desire to act as a free, autonomous moral agent that makes these human minds resist accepting such responsibility.

This resistance or opposition may come from the car manufacturing companies who want to strike the ego of individuals with misleading commercials promising them stress-free rides through busy traffic while fully relying on technology. It may also come from individuals who, by such misleading propaganda, turn this technology into an idol, which has to serve the modern man’s egocentric view of a free and autonomous agent. As indicated on page 63, Bonhoeffer has called this a false idol that may lead to a habitual obsession that hampers being human.

In line with this thinking, such opposition may endanger the realization of the promised benefits of AI for ADS in making driving safer, reducing pollution, and increasing the flexibility of the disabled. It is here that I foresee still an important role for the theologian in changing the mindset of modern Western people. As this change in anthropology is closely related to Niebuhr’s view on human conscience, we should start by putting Christian theology again in the forefront when analyzing ethical concerns and questions concerning new technological developments.

In Section VIII.5, I used the insights of Bonhoeffer and Niebuhr to overcome this resistance. It calls for an orientation of the self towards the other, thereby avoiding the two dangers inherent in living a vicarious life, as indicated by Bonhoeffer in Section IV.3.3, namely by absolutizing the self or absolutizing the other. Falling into these pitfalls turns responsibility into a self-made abstract idol.

This re-orientation requires a social and personal reinterpretation of the remembered past and anticipated future. A radical change is necessary to make such reinterpretation possible, and to make this radical change, Niebuhr asserts the following:

> Yet all of these social and personal reinterpretations of remembered past and anticipated futures do not radically change either our general pattern of understanding of action upon us or our general mode of fitting response so long as our sense of the ultimate context remains unrevised.[373]

Christians in general and Christian theologians, in particular, can make an important contribution to this revision, namely, the communication of as well as embodying Jesus' commandment to "[l]ove your neighbor as yourself" (Matthew 22:39 - NIV). Moreover, this should start with loving "the Lord your God with all your heart and with all your soul and with all your mind" (Matthew 22:37 - NIV). By realizing this last statement, a change of mindset can begin. The contribution of Bonhoeffer and Niebuhr – to embody what they taught – can be an even more significant source of inspiration for the Christian and Christian theologian than their theological writings are.

[373] Niebuhr, 106.

Bibliography

Asaro, Peter. "On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making." *International Review of the Red Cross* 94, no. 886 (June 2012): 687–709. https://doi.org/10.1017/S1816383112000768.

Austad, Torleiv. "'The Responsibility of the Church for Society' and Other Essays by H. Richard Niebuhr." *European Journal of Theology* 21, no. 1 (April 2012): 71–72.

Aydin, Ciano, Margoth González Woge, and Peter-Paul Verbeek. "Technological Environmentality: Conceptualizing Technology as a Mediating Milieu." *Philosophy & Technology* 32, no. 2 (June 1, 2019): 321–38. https://doi.org/10.1007/s13347-018-0309-3.

Barnett, John. "Training People to Use Automation: Strategies and Methods." *Journal of Systemics, Cybernetics and Informatics* 3, no. 5 (October 1, 2005): 73–76.

Bhargava, Vikram, and Tae Wan Kim. "Autonomous Vehicles and Moral Uncertainty." In *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, edited by Patrick Lin, Keith Abney, and Ryan Jenkins, 5–19. Oxford, New York, NY: Oxford University Press, 2017. https://doi.org/10.1093/oso/97880190652951.001.0001.

Boddington, Paula. *Towards a Code of Ethics for Artificial Intelligence*. Artificial Intelligence: Foundations, Theory, and Algorithms. New York, NY: Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-60648-4.

Bonhoeffer, Dietrich. *Discipleship*. Minneapolis, MN: Fortress Press, 2015.

Bonhoeffer, Dietrich. *Ethics*. Edited by Clifford J. Green. Translated by Reinhard Krauss, Charles C. West, and Douglas W. Stott. Dietrich Bonhoeffer Works. Vol. 6. Minneapolis, MN: Fortress Press, 2005.

Bonhoeffer, Dietrich. *Letters and Papers from Prison*. Minneapolis, MN: Fortress Press, 2010.

Bonhoeffer, Dietrich. *Sanctorum Communio: A Theological Study of the Sociology of the Church*. Edited by Clifford J. Green. Translated by Joachim Von Soosten, Reinhard Kraus, and Nancy Lukens. Annotated edition. Minneapolis, MN: Fortress Press, 2009.

Bonhoeffer, Dietrich. *Widerstand und Ergebung*. Edited by Christian Gremmels, Eberhard Bethge, and Renate Berthge. Dietrich Bonhoeffer Werke. Vol. 8. Gütersloh: Gütersloher Verlagshaus, 1998.

Bratman, Michael E. "Two Faces of Intention." *Philosophical Review* 93, no. 3 (July 1984): 375–405. https://doi.org/10.2307/2184542.

Bratman, Michael E. “Fischer and Ravizza on Moral Responsibility and History.” *Philosophy and Phenomenological Research* 61, no. 2 (2000): 453–58. https://doi.org/ppr2000612106.

Brooks, Rodney A. “Intelligence without Reason.” In *Proceedings of the 12th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'91, San Francisco, CA: Morgan Kaufmann Publishers, 1991, 569–95.

Brunner, H. E. *Truth As Encounter*. Philadelphia, PA: Westminster John Knox Press, 2000.

Daniel, Joshua. “H. Richard Niebuhr’s Reading of George Herbert Mead: Correcting, Completing, and Looking Ahead.” *Journal of Religious Ethics* 44, no. 1 (February 18, 2016): 92–115. https://doi.org/10.1111/jore.12133.

Daniel, Joshua. *Transforming Faith: Individual and Community in H. Richard Niebuhr*. Eugene, ORE: Pickwick Publications, 2015.

Dastin, Jeffrey. “Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women.” *Reuters*, October 10, 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

Dennett, Daniel C. *Elbow Room: The Varieties of Free Will Worth Wanting*. NED-New edition. Cambridge: The MIT Press, 2015. http://www.jstor.org/stable/j.ctt17kk7ns.

Dietrich, Manuel, and Thomas H. Weisswange. “Distributive Justice as an Ethical Principle for Autonomous Vehicle Behavior beyond Hazard Scenarios.” *Ethics and Information Technology* 21, no. 3 (September 1, 2019): 227–39. https://doi.org/10.1007/s10676-019-09504-3.

Eliot, Lance. “Has Elon Musk Set Up Regulatory Boogeyman As Scapegoat For Ongoing Delay In Promise Of Self-Driving Teslas?” *Forbes*, April 17, 2020. https://www.forbes.com/sites/lanceeliot/2020/04/17/has-elon-musk-set-up-a-regulatory-boogeyman-as-scapegoat-for-delay-in-his-self-driving-tesla-promise/.

Elliott, Carl. “Enhancement Technologies and the Modern Self.” *Journal of Medicine and Philosophy* 36, no. 4 (August 2011): 364–74. https://doi.org/10.1093/jmp/jhr031.

Encyclopedia Britannica. “Computer Science - Algorithms and Complexity.” Accessed March 23, 2021. https://www.britannica.com/science/computer-science.

Fischer, John Martin, and Mark Ravizza. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press, 1998.

Foot, Philippa. “The Problem of Abortion and the Doctrine of the Double Effect.” *Oxford Review* 5 (1967): 5–15.

Frankfurt, Harry G. “Freedom of the Will and the Concept of a Person.” *Journal of Philosophy* 68, no. 1 (January 1971): 5–20. https://doi.org/10.2307/2024717.

Frick, Peter. “Dietrich Bonhoeffer: Engaging Intellect – Legendary Life.” *Religion Compass* 6, no. 6 (June 28, 2012): 309–22. https://doi.org/10.1111/j.1749-8171.2012.00357.x.

Fritz, Alexis, Wiebke Brandt, Henner Gimpel, and Sarah Bayer. “Moral Agency without Responsibility? Analysis of Three Ethical Models of Human-Computer Interaction in Times of Artificial Intelligence (AI).” *De Ethica* 6, no. 1 (June 30, 2020): 3–22. https://doi.org/10.3384/de-ethica.2001-8819.20613.

Fujiyoshi, Hironobu, Tsubasa Hirakawa, and Takayoshi Yamashita. “Deep Learning-Based Image Recognition for Autonomous Driving.” *IATSS Research* 43, no. 4 (December 1, 2019): 244–52. https://doi.org/10.1016/j.iatssr.2019.11.008.

Geiger, R. Stuart, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. “Garbage In, Garbage Out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From?” In *Proceedings of the 2020 Conference on Fairness, Accountability and Transparency*, Barcelona, December 17, 2019, 325–336. https://doi.org/10.1145/3351095.3372862.

Gurney, Jeffrey. “Crashing into the Unknown: An Examination of Crash-Optimization Algorithms Through the Two Lanes of Ethics and Law.” *Albany Law Review* 79, no. 183 (March 2016). https://papers.ssrn.com/abstract=2622125.

Gustafson, James W. *Introduction of The Responsible Self: An Essay in Christian Moral Philosophy*. New York, NY: Harper & Row, 1963.

Harari, Yuval Noah. *Homo Deus: A Brief History of Tomorrow*. Illustrated edition. New York, NY: Harper & Row, 2017.

Heuvel, Steven C. van den. *Bonhoeffer’s Christocentric Theology and Fundamental Debates in Environmental Ethics*. Eugene, ORE: Pickwick Publications, an Imprint of Wipf and Stock Publisher, 2017.

Heuvel, Steven C. van den. “Leadership and the Ethics of Responsibility: An Engagement with Dietrich Bonhoeffer.” In *The Challenges of Moral Leadership*, edited by Patrick Nullens and Steven C. van den Heuvel, 111–25. Christian Perspectives on Leadership and Social Ethics 2. Leuven: Peeters, 2016.

Horowitz, Michael C., and Paul Scharre. *Meaningful Human Control in Weapon Systems: A Primer*. Washington D.C.: Center for a New American Security, 2015. https://www.jstor.com/stable/resrep06179.

Huber, Wolfgang. “Ethics of Responsibility in a Theological Perspective.” *Stellenbosch Theological Journal* 6, no. 1 (August 2020): 185–206. https://doi.org/10.17570/stj.2020.v6n1.a11.

Huber, Wolfgang. “Toward an Ethics of Responsibility.” *Journal of Religion* 73, no. 4 (October 1, 1993): 573–91. https://doi.org/10.1086/489259.

Jong, Roos de. “The Retribution-Gap and Responsibility-Loci Related to Robots and Automated Technologies: A Reply to Nyholm.” *Science and Engineering Ethics* 26, no. 2 (April 1, 2020): 727–35. https://doi.org/10.1007/s11948-019-00120-4.

Kant, Immanuel. *Groundwork for the Metaphysics of Morals*. Edited by Thomas E. Hill Jr and Arnulf Zweig. Oxford Philosophical Texts. Oxford, New York, NY: Oxford University Press, 2002.

Keeling, Geoff. “Why Trolley Problems Matter for the Ethics of Automated Vehicles.” *Science and Engineering Ethics* 26, no. 1 (February 1, 2020): 293–307. https://doi.org/10.1007/s11948-019-00096-1.

Kochenderfer, Mykel J., Christopher Amato, Girish Chowdhary, Jonathan P. How, and Hayley J. Davison Reynolds. *Decision Making Under Uncertainty: Theory and Application*. Illustrated edition. Cambridge: The MIT Press, 2015.

Lehmann, Paul. *Ethics in a Christian Context*. New York, NY: Harper & Row, 1963.

Lemay, Mathieu. “Understanding Canada’s Algorithmic Impact Assessment Tool.” *Medium*, June 11, 2019. https://towardsdatascience.com/understanding-canadas-algorithmic-impact-assessment-tool-cd0d3c8cafab.

Liddell, H. G., R. Scott, Sir Henry Stuart Jones, Roderick McKenzie, P. G. W. Glare, and A. A. Thompson. *A Greek-English Lexicon*. 9th edition with Revised Supplement. Oxford, New York, NY: Oxford University Press, 1996.

Lin, Patrick, Keith Abney, and Ryan Jenkins, eds. *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. Oxford, New York, NY: Oxford University Press, 2017.

Matthias, Andreas. “The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata.” *Ethics and Information Technology* 6, no. 3 (September 1, 2004): 175–83. https://doi.org/10.1007/s10676-004-3422-1.

Mecacci, Giulio, and Filippo Santoni de Sio. “Meaningful Human Control as Reason-Responsiveness: The Case of Dual-Mode Vehicles.” *Ethics and Information Technology* 22, no. 2 (June 1, 2020): 103–15. https://doi.org/10.1007/s10676-019-09519-w.

Meyer, Gereon, and Sven Beiker, eds. *Road Vehicle Automation 6*. Lecture Notes in Mobility. New York, NY: Springer International Publishing, 2019. https://doi.org/10.1007/978-3-030-22933-7.

Michon, John A. "A Critical View of Driver Behavior Models: What Do We Know, What Should We Do?" In *Human Behavior and Traffic Safety*, edited by Leonard Evans and Richard C. Schwing, 485–524. Boston, MA: Springer US, 1985. https://doi.org/10.1007/978-1-4613-2173-6_19.

Millar, Jason. "Ethics Setting for Autonomous Vehicles." In *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, edited by Patrick Lin, Abney Keith, and Jenkins Ryan, 20–34. Oxford, New York, NY: Oxford University Press, 2017.

Moral Machine. "Moral Machine." Accessed June 30, 2020. http://moralmachine.mit.edu.

Morgan, Jeffrey. "A Loss of Judgment: The Dismissal of the Judicial Conscience in Recent Christian Ethics." *Journal of Religious Ethics* 45, no. 3 (August 14, 2017): 539–61. https://doi.org/10.1111/jore.12189.

Nat'l Highway Safety Admin., U.S. Dep't of Transp. *Countermeasures That Work: A Highway Safety Countermeasure Guide for State Highway Safety Offices*. Countermeasures that Work. Chapel Hill, NA: University of North Carolina, 2011. http://www.nhtsa.gov/staticfiles/nti/pdf/811444.pdf.

Niebuhr, H. Richard. *Radical Monotheism and Western Culture, With Supplementary Essays*. Louisville, KY: Westminster John Knox Press. 1993.

Niebuhr, H. Richard. "The Ego-Alter Dialectic and the Conscience." *Journal of Philosophy* 42, no. 13 (June 1945): 352–59. https://doi.org/10.2307/2018981.

Niebuhr, H. Richard. *The Responsible Self: An Essay in Christian Moral Philosophy*. New York, NY: Harper & Row, 1963.

Nullens, Patrick, and Ronald T. Michener. *The Matrix of Christian Ethics: Integrating Philosophy and Moral Theology in a Postmodern Context*. London: IVP Books, 2010.

Nyholm, Sven. "Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci." *Science and Engineering Ethics* 24, no. 4 (August 1, 2018): 1201–19. https://doi.org/10.1007/s11948-017-9943-x.

Nyholm, Sven, and Jilles Smids. "The Ethics of Accident-Algorithms for Self-Driving Cars: An Applied Trolley Problem?" *Ethical Theory and Moral Practice* 19, no. 5 (November 1, 2016): 1275–89. https://doi.org/10.1007/s10677-016-9745-2.

O'Connell, Mary Ellen. "Chapter 12. Banning Autonomous Killing: The Legal and Ethical Requirement That Humans Make Near-Time Lethal Decisions." In *The American Way of Bombing Changing Ethical and Legal Norms, from Flying Fortresses to Drones*, edited by M. Evangelista and H. Shue, 224–36. Ithaca, NY: Cornell University Press, 2018. https://doi.org/10.7591/9780801454578-014.

Paeth, Scott. "The Responsibility to Lie and the Obligation to Report." *Journal of Business Ethics* 112, no. 4 (January 2013): 559–66.

Pfeifer, Rolf, and Fumiya Iida. "Embodied Artificial Intelligence: Trends and Challenges." In *Embodied Artificial Intelligence: International Seminar, Dagstuhl Castle, Germany, July 7-11, 2003. Revised Papers*, edited by Fumiya Iida, Rolf Pfeifer, Luc Steels, and Yasuo Kuniyoshi, 1–26. Lecture Notes in Computer Science. Berlin: Springer Gabler, 2004. https://doi.org/10.1007/978-3-540-27833-7_1.

Premio Inc. "CANBus: The Central Networking System of Vehicles." June 20, 2019. https://premioinc.com/blogs/blog/can-bus-the-central-networking-system-of-vehicles.

Puffer, Matthew. "Three Rival Versions of Moral Reasoning: Interpreting Bonhoeffer's Ethics of Lying, Guilt, and Responsibility." *Harvard Theological Review* 112, no. 2 (April 2019): 160–83. https://doi.org/10.1017/S001781601900004X.

Quach, Katyanna. "Remember the Uber Self-Driving Car That Killed a Woman Crossing the Street? The AI Had No Clue about Jaywalkers." *The Register*. November 6, 2019. https://www.theregister.com/2019/11/06/uber_self_driving_car_death/.

Rawlings, J. B. "Tutorial Overview of Model Predictive Control." *IEEE Control Systems Magazine* 20, no. 3 (June 2000): 38–52. https://doi.org/10.1109/37.845037.

Rawls, John. *A Theory of Justice. Revised edition*. Cambridge, MA: Harvard University Press, 2009. https://www.hup.harvard.edu/catalog.php?isbn=9780674000780.

Reed, Esther D. "The Limits of Individual Responsibility: Dietrich Bonhoeffer's Reversal of Agent-Act-Consequence." *Journal of the Society of Christian Ethics* 37, no. 2 (Fall/Winter 2017): 39–58. https://www.jstor.org/stable/44987550

Renda, Andrea. "Artificial Intelligence: Ethics, Governance and Policy Challenges." *CEPS Task Force*, February 15, 2019. https://www.ceps.eu/ceps-publications/artificial-intelligence-ethics-governance-and-policy-challenges/.

Rios, Jeremy M. "Bonhoeffer and Bowen Theory: A Theological Anthropology of the Collective-Person and Its Implications for Spiritual Formation." *Journal of Spiritual Formation and Soul Care* 13, no. 2 (November 1, 2020): 176–92. https://doi.org/10.1177/1939790920915700.

SAE. “Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles.” *Technical*, June 15, 2018. https://www.sae.org/standards/content/j3016_201806/.

Saleh, Naveed. “The Dangers of Alarm Fatigue.” *Psychology Today*, January 12, 2018. https://www.psychologytoday.com/blog/the-red-light-district/201801/the-dangers-alarm-fatigue.

Santoni de Sio, Filippo, and Jeroen van den Hoven. “Meaningful Human Control over Autonomous Systems: A Philosophical Account.” *Frontiers in Robotics and AI* 5 (February 18, 2018). https://doi.org/10.3389/frobt.2018.00015.

Schweiker, William. *Responsibility and Christian Ethics*. Cambridge: Cambridge University Press, 1999.

Sedgwick, Timothy F. “Niebuhr’s Ethic of Responsibility: A Unified Interpretation.” *Saint Luke’s Journal of Theology* 23, no. 4 (September 1980): 265–83.

Shepardson, David. “Tesla, NTSB Clash over Autopilot Investigation.” *Reuters*, April 12, 2018. https://www.reuters.com/article/us-tesla-crash-autopilot-idUSKBN1HJ2JS.

Smuha, Nathalie. “Ethics Guidelines for Trustworthy AI.” *European Commission*, April 8, 2019. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Sun, S., Z. Cao, H. Zhu, and J. Zhao. “A Survey of Optimization Methods From a Machine Learning Perspective.” *IEEE Transactions on Cybernetics* 50, no. 8 (August 2020): 3668–81. https://doi.org/10.1109/TCYB.2019.2950779.

Taylor, Charles. *The Malaise of Modernity*. Concord, Canada: Anansi Press, 1991. https://www.goodreads.com/work/best_book/87699794-the-malaise-of-modernity.

Thornton, Sarah M., Benjamin Limonchik, Francis E. Lewis, Mykel J. Kochenderfer, and J. Christian Gerdes. “Toward Closing the Loop on Human Values.” *IEEE Transactions on Intelligent Vehicles* 4, no. 3 (September 2019): 437–46. https://doi.org/10.1109/TIV.2019.2919471.

Thornton, Sarah M., Selina Pan, Stephen M. Erlien, and J. Christian Gerdes. “Incorporating Ethical Considerations Into Automated Vehicle Control.” *IEEE Transactions on Intelligent Transportation Systems* 18, no. 6 (June 2017): 1429–39. https://doi.org/10.1109/TITS.2016.2609339.

Thornton, Sarah Marie. “Autonomous Vehicle Motion Planning with Ethical Considerations.” PhD diss., University of Stanford, Stanford, 2018. https://searchworks.stanford.edu/view/12746436.

Tillich, Paul. “The Method of Correlation,” In ibid., Systematic Theology., 1:59–66. Chicago, ILL: University of Chicago Press, 1950.

UK Government. “Killer Robots: UK Government Policy on Fully Autonomous Weapons.” April 2013. http://www.article36.org/wp-content/uploads/2013/04/Policy_Paper1.pdf.

Vallor, Shannon. *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford, New York, NY: Oxford University Press, 2016.

Verbeek, Peter-Paul. “Some Misunderstandings About the Moral Significance of Technology.” In *The Moral Status of Technical Artefacts*, edited by Peter Kroes and Peter-Paul Verbeek, 75–88. Philosophy of Engineering and Technology. Dordrecht: Springer Netherlands, 2014. https://doi.org/10.1007/978-94-007-7914-3_5.

Vignard, Kerstin. “The Weaponization of Increasingly Autonomous Technologies: Considering How Meaningful Human Control Might Move Discussion Forward.” *UNIDIR*, 2014. https://www.unidir.org/files/publications/pdfs/considering-how-meaningful-human-control-might-move-the-discussion-forward-en-615.pdf.

Vincent, James. “Global Preferences for Who to Save in Self-Driving Car Crashes Revealed.” *The Verge*, October 24, 2018. https://www.theverge.com/2018/10/24/18013392/self-driving-car-ethics-dilemma-mit-study-moral-machine-results.

Volvo Cars. “Autonomous Driving | Intellisafe | Volvo Cars.” Accessed June 26, 2020. https://www.volvocars.com/en-kw/own/own-and-enjoy/autonomous-driving.

Wallach, Wendell. “Robot Minds and Human Ethics: The Need for a Comprehensive Model of Moral Decision Making.” *Ethics and Information Technology* 12, no. 3 (July 2010): 243–50. https://doi.org/10.1007/s10676-010-9232-8.

Weber, Max, and Ronald Speirs. “Political Writings.” Cambridge: Cambridge University Press, 1994. https://doi.org/10.1017/CBO9780511841095.

Yardon, Danny, and Dan Tynan. “Tesla Driver Dies in First Fatal Crash While Using Autopilot Mode.” *Guardian*. June 30, 2016. http://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk.

Yurtsever, Ekim, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. “A Survey of Autonomous Driving: Common Practices and Emerging Technologies.” *IEEE Access* 8 (2020): 58443–69. https://doi.org/10.1109/ACCESS.2020.2983149.

Zhang, Xiang, Wei Liu, and S. Travis Waller. “A Network Traffic Assignment Model for Autonomous Vehicles with Parking Choices.” *Computer-Aided Civil and Infrastructure Engineering* 34, no. 12 (July 30, 2019): 1100–1118. https://doi.org/10.1111/mice.12486.

Zimmermann, Jens. “Virtue Ethics and Realistic Responsibility in an Age of Globalization.” In *Handbook of Virtue Ethics in Business and Management*, edited by Alejo José G. Sison, Gregory R. Beabout, and Ignacio Ferrero, 1281–95. International Handbooks in Business Ethics. Dordrecht: Springer Netherlands, 2017. https://doi.org/10.1007/978-94-007-6510-8_54.