Towards trustworthy intelligent vehicle technology development

Kaspar Raats
“With great power there must also come – great responsibility!”

Stan Lee, Amazing Fantasy No. 15
This thesis addresses the unresolved issues of responsibility and accountability in autonomous vehicle (AV) development, advocating for human-centred approaches to enhance trustworthiness. While AVs hold the potential for improved safety, mobility, and environmental impact, poorly designed algorithms pose risks, leading to public distrust. Trust research focuses on technology-related aspects but overlooks trust within broader social and cultural contexts. Efforts are underway to understand algorithm design practices, acknowledging their potential unintended consequences. For example, Baumer (2017) advocates human-centred algorithm design (HCAD) to align with user perspectives and reduce risks. HCAD incorporates theoretical, participatory, and speculative approaches, emphasising user and stakeholder engagement. This aligns with broader calls for prioritising societal considerations in technology development (Stilgoe, 2013). The research in this thesis responds to these calls by integrating theories on trust and trustworthiness, autonomous vehicle development, and human-centred approaches in empirical investigations guided by the following research question: “How can human-centred approaches support the development of trustworthy intelligent vehicle technology?” This thesis approaches the question through design ethnography to ground the explorations in people’s real-life routines, practices and anticipations and demonstrate how design ethnographic techniques can infuse AV development with human-centred understandings of people’s trust in AVs. The studies reported in this thesis include a) interviews and participatory observations of algorithm designers, b) interviews and probing with residents, and c) staging collaborative, reflective practice through the design ethnographic materials and co-creation with citizens, city, academic and industry stakeholders, including AV algorithm designers.

Through these empirical explorations, this thesis suggests an answer to the research question by coining a novel and timely framework for intelligent vehicle development: trustworthy algorithm design (TAD). TAD demonstrates trustworthiness as an ongoing process, not just a measurable outcome from human-technology interactions. It calls to consider autonomous vehicle algorithms as construed through a network of stakeholders, practices, and technologies and, therefore, defines trustworthy algorithm design as a continuous collaborative learning and evolvement process of different disciplines and sectors. Furthermore, the TAD framework suggests that for autonomous vehicle algo-
rithm design to be trustworthy, it must be responsive, interventional, intentional and transdisciplinary.

The TAD framework integrates ideas and strategies from different well-known trajectories of research in the field of responsible and human-centred technology development: Human-Centred Algorithm Design (Baumer, 2017), algorithms as culture (Seaver, 2017) and Responsible Innovation (Stilgoe et al., 2013). The thesis contributes to this field by empirically investigating how this integrated framework helps expand existing understandings of interactional trust in intelligent technologies and include the relevance of participatory processes of trustworthiness and how these processes are nurtured through cross-sector co-learning and design ethnographic materials.
Acknowledgements

It is the people (and dogs) you meet, travel with, experience things together, learn from, bring joy to, get upset at, laugh and cry with who make or break a good PhD journey. I have had the pleasure of meeting many wonderful, intelligent, and smart people, some already before the PhD started, who have influenced my current journey.

Thank you, Vaike Fors. I cannot imagine having had a better main supervisor. You orchestrated a truly amazing experience for me. Your guidance, support, and caring, perfectly timed, pedagogically excellent educational engagements were essential to make this experience something to miss. Thank you, Magnus Bergquist and Jesper Lund. As co-supervisors, your input was always valuable, on point, and appreciated. It truly elevated my research. Thank you, Robert Broström. You were like a ninja. Your support as my industrial supervisor was barely noticeable until I realised that the uninterrupted time to focus on my research was primarily because of you. Thank you, Sarah Pink. Your guidance and support have been invaluable. I have truly enjoyed learning and collaborating with you.

Thank you, Susanne Duvall Innings, Suzanne Andersson, and Patrik Palo. Your role was essential for the access to Helsingborg City, Gothenburg City and Volvo Cars’s stakeholders. Your enthusiasm, trust and support in such an explorative research project like AHA was admirable. Similarly, thank you, Azra Habibovic, Alexander Eriksson, David Friberg, Jonas Fenn, and Jens Lundström, for gaining me entry to the wonderful world of algorithm design and algorithm designers. Furthermore, thank you to all the research participants, project stakeholders, residents and algorithm designers who participated in different research activities. Your time, input, and trust were vital in developing this thesis.

Thank you, Thomas Lindgren. We ate so many breakfasts together that I could still taste the oat porridge with fresh papaya in my mouth. I agree with you that one of the best parts of this thesis has been our conversations, especially during the breakfasts in Melbourne and Kavala. I look forward to continuing our discussions and collaboration back in the industry.

Thank you, Katalin Osz, Meike Brodersen, and Esbjörn Ebbesson. Collaborating, travelling, and hanging out with you have been my pleasure and one of the highlights of this journey. Your intelligence, knowledge, and capability humble me. You are the ones I want to become when I grow up. Furthermore,
thank you to all the other wonderful colleagues at E4 who have contributed time and effort to improve or support this research. Special thank you to Maria Åkesson and Michel Thomsen. You have been part of this journey since the beginning and have always provided constructive criticism and guidance. I truly appreciate it.

As an industrial PhD, many industry colleagues have also participated in this experience. Therefore, thank you, Jon Seneger. The first time we met, I remember you saying, “If you want to do something interesting, contact me.” I have truly enjoyed our conversations and collaborations before and during the PhD and appreciate you for all the people and opportunities you have introduced me to. I believe our different endeavours have also led to this PhD opportunity. Thank you, Lingxi Yuan, Erick Hernández, and Aljoscha Ledwa. You helped me visualise research and exhibition materials that had an essential role in the studies and other AHA project activities and will most likely live beyond this PhD. You can be trusted to do an excellent job and be in good spirits.

Thank you, Timotei Florin Ghiurau, Henrik Wiberg, Martin Ivarson, Jakob Widerberg, and Henrik Harplinger. Your perspectives, ability to connect dots, willingness to learn and share knowledge, and sense of humour are remarkable and uplifting. Our Friday meeting has been the best and most anticipated part of my industry work for years. Furthermore, you have provided me with knowledge and opportunities I would have never imagined.

Thank you, Hannaneh Yazdi, Eleanor Chin Derix, Erika Renado, Fjollë Novakazi, and Yuchu Liu. As fellow travellers, you have made this long journey much more exciting and fun. Your intelligence, capability, and resilience inspire me. You are truly unique.

As the saying goes, a healthy mind in a healthy body. Thank you, Carl Branting, for urging me to go climbing with you, especially when I did not feel like doing anything. It has been filled with great and occasionally scary moments and helped me maintain a clear and calm head.

Last but not least, I thank my family. Thank you, Mother, for the unconditional support and caring. You always worry if I have enough time to focus, stay on track, and eat properly. I appreciate it. Thank you, Anna; you are my biggest critic; Klaus, my greatest fan; and Jaan, you were very helpful while Mom and I simultaneously pursued our PhDs. You are all brilliant, intelligent, funny, and capable people, and you make me want to be at least a tad better. Thank you, Pille, for enduring my being away for long periods, constantly discussing the thesis, or occupying your time rehearsing my presentation. We have had a slightly different pace than usual in the past few years, but we have finally made it. #youfirst

You all made this PhD a great experience, and with all of you, I would do it again in a heartbeat. Thank you. Love you all!
Before going into this thesis, a brief overview of the personal and project context of the work. I was already employed in the automotive industry when embarking on this research. This meant splitting time between industry and academic work. Working in the industry had implications for this thesis’s study context. Autonomous vehicles were still one of the prevalent topics in the automotive industry, and there was a strong interest in understanding how to induce user and public trust in these vehicles. My interest, however, was triggered by design ethnography as the underlying methodological approach in this project. My interest towards design ethnography had already been sparked during my master’s studies around 2012 when we were presented with projects utilising design ethnography in urban development and industrial development projects. This curiosity was deepened when I was offered an opportunity to partake in a project around an urban cycling ecosystem involving stakeholders from industry, design consultancy and research institute while carrying out my master’s thesis research. Luckily, after many years of mainly evaluative work in the industry, I could finally engage in my own design ethnographic research project.

This PhD research was conducted within three VINNOVA-funded research projects - TIC (reference nr. 2017-03058), AHA (reference nr. 2018-02088) and AHA 2 (reference nr. 2019-04786). These projects also affected the study contexts of this thesis. Trust in Intelligent Cars (TIC) project aimed at combining experimental car studies with an ethnographic approach to investigate user experience of automated driving and the impact increased intelligence in and around vehicles has on people’s trust in AVs. AHA continued similarly and aimed to develop a method combining experimental testing with ethnography to design intelligent services and urban planning strategies. AHA 2 took it a step further and, through Urban Living Lab methodology, focused on developing a method involving users and public and private stakeholders in co-creating future mobility. These projects (hereafter referred to as AHA) enabled me to experiment with design ethnography, design fiction, value co-creation, responsible innovation, and reflective practice, go in-depth with concepts of trust and trustworthiness and explore AV algorithm design practice. All of which I sincerely appreciate.

Striving to develop autonomous vehicles is a complex topic. In the long run, they are anticipated to bring societal benefits like increased traffic safety and
decreased environmental impact. However, there does not seem to be a clear understanding of what tangible value they can offer to people’s daily lives. For the sake of this thesis, I assume that autonomous vehicles can bring the envisioned long-term benefits and explore how human-centred approaches can support the development to become more responsible and trustworthy.
Appended Publications

PAPER 1

Contribution: Raats was principal author and planned, carried out and analysed the research, with support in analysis and writing from Fors and writing input from Pink

PAPER 2

Contribution: Raats was principal author and planned, carried out and analysed the research, with support in analysis and writing from Lund, and support in carrying out research and writing from Brodersen

PAPER 3

Contribution: Raats was principal author and planned, carried out and analysed the research, with support in carrying out research, analysis and writing from Bergquist and writing input from Fors

PAPER 4
Contribution: Raats was principal author and planned, carried out and analysed the research, with support in analysis and writing from Fors and support in carrying out research from Ebbesson

PAPER 5
Additional Publications

Papers where my research contributes and that illustrate my research journey but are not appended:


## Contents

Introduction ................................................................................................................. 17  
Aim and Research Question .................................................................................... 20  
Thesis Outline ........................................................................................................... 21  
Key Concepts ........................................................................................................... 21  

Related Theory and Research .................................................................................. 25  
  Human-Centred Design of Autonomous Vehicles .............................................. 25  
  Human-Centred Autonomous Vehicle Design Approaches ................................. 27  
Trust in Autonomous Vehicles ............................................................................. 31  
  Defining trustworthiness ....................................................................................... 38  
Towards Understanding Trustworthy Algorithm Design ....................................... 41  

Research Methodology ............................................................................................ 45  
  Design Ethnography ............................................................................................... 45  
  Data Production ..................................................................................................... 46  
  Data Analysis .......................................................................................................... 52  
  Research Quality and Ethics .................................................................................. 53  

Summary of Papers .................................................................................................. 57  
  Paper 1 ................................................................................................................... 58  
  Paper 2 ................................................................................................................... 60  
  Paper 3 ................................................................................................................... 62  
  Paper 4 ................................................................................................................... 64  
  Paper 5 ................................................................................................................... 67  

Discussion .................................................................................................................. 71  
  Trustworthy Algorithm Design ............................................................................. 72  
    TAD Element 1: Responsive ............................................................................... 73  
    TAD Element 2: Interventional .......................................................................... 74  
    TAD Element 3: Intentional .............................................................................. 76  
    TAD Element 4: Transdisciplinary ................................................................... 77  
  Contribution and Implications ............................................................................. 79  
  Limitations and Future Research ......................................................................... 81  

Concluding Remarks ................................................................................................... 83  

References .................................................................................................................. 85
# Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI</td>
<td>Artificial Intelligence</td>
</tr>
<tr>
<td>AV</td>
<td>Autonomous Vehicle</td>
</tr>
<tr>
<td>DE</td>
<td>Design Ethnography</td>
</tr>
<tr>
<td>HCAD</td>
<td>Human-Centred Algorithm Design</td>
</tr>
<tr>
<td>HCD</td>
<td>Human-Centred Design</td>
</tr>
<tr>
<td>HCI</td>
<td>Human-Computer Interaction</td>
</tr>
<tr>
<td>RI</td>
<td>Responsible Innovation</td>
</tr>
<tr>
<td>TAD</td>
<td>Trustworthy Algorithm Design</td>
</tr>
<tr>
<td>UCD</td>
<td>User-Centred Design</td>
</tr>
</tbody>
</table>
Introduction

Will autonomous vehicles (AVs) ever be trustworthy? Perhaps not. Will people ever trust AVs? Probably. However, the distinction between trust and trustworthiness lies in who takes the trust responsibility. Through several incidents with AVs that have crashed and even killed pedestrians, it has become evident that questions about responsibility and accountability in intelligent vehicle development are yet to be solved. On this broader scope, this thesis aims to contribute with research-based suggestions and ideas for how human-centred approaches can support trustworthy intelligent vehicle technologies.

At the heart of AVs and other intelligent vehicles lie algorithms, which are "formalised descriptions of a computational procedure" (Dourish, 2006, p. 3) that instruct the computational systems in the AVs on what to do and how to operate independently. It is predicted that if the development of fully self-driving vehicles succeeds, it will increase traffic safety, provide mobility for non-drivers, and decrease the environmental impact of cars (Litman, 2021). On the other hand, as driving is a highly complex task, unsuccessfully designed algorithms can result in fatalities and public rejection of AVs. For example, the American Automotive Association reported that only 12% of drivers anticipated trusting a self-driving vehicle (AAA Public Affairs, 2020), and only 26% thought AVs were a good idea for society (Maslej et al., 2023). This is problematic, as public trust is recognised as essential for people to accept innovative technologies (Lüders et al., 2017) and key to the success of intelligent technologies like AVs (Wintersberger & Riener, 2016). According to these calls, trust in AVs must increase if society is to reap the anticipated benefits of these technologies. This debate has caused recent efforts to comprehend algorithm design practice.

Algorithms, shaped by the values and assumptions of their designers, can inadvertently lead to real-world harm. Yet, designers may encounter difficulty articulating the precise outcomes generated by the algorithm. In response to this problem, Baumer (2017) advocates for human-centred strategies in algorithm design to mitigate risks and enhance alignment with users' perceptions. This approach, called the human-centred algorithm design (HCAD) framework, incorporates theoretical, participatory, and speculative
strategies. HCAD encourages designers to engage users and stakeholders, ensuring algorithms address the correct problems and anticipate potential societal impacts. Baumer’s framework has influenced subsequent algorithm design efforts (e.g., Cherrington et al., 2020; Dekker et al., 2022; Flügge et al., 2020), emphasising a deep understanding of people’s daily practices and relationships. These strategies echo broader calls for prioritising the consideration of people and society in technology development. The research presented in this thesis is done in response to this call for HCAD by developing and demonstrating methodologies that bring together research on trust, autonomous vehicle development, and human-centred approaches to support what is coined in this thesis as trustworthy algorithm design. By creating this framework, the thesis will also contribute to a deeper understanding of how human-centred approaches can develop the theory and practice of researching trust in autonomous vehicles by reframing the concept of trust through the lens of trustworthiness, as described below.

In the wake of the hype caused by the rapid development and testing of AVs on public roads, a significant investment in Human-Computer Interaction (HCI) research on how to make people trust these vehicles has followed. Previous research has assumed trust to develop in momentary and individual interactions between a person and technology and to be transactional. Trust has mainly been studied in laboratory experiments and self-assessed by the participants through questionnaires, while the focus has commonly been on understanding how system performance and design features influence trust in AV development (Raats et al., 2020). However, in 1994, Bonnie M. Muir pointed out the necessitating exploration of trust in AVs beyond the technology itself, encompassing a broader network of trust, including developers, designers, city infrastructure architects, and lawmakers (Muir, 1994). Despite this, the development of more technology-driven understandings of how to make people accept and adopt technology by altering technological factors that make the technology feel reliable and easy to use (Davis, 1989; Taylor & Todd, 1995; Venkatesh et al., 2003) seems to be prevailing. Despite this bulk of research offering a significant amount of knowledge on how trust develops when people interact with automation and AVs, the empirical studies have, at the same time, neglected to explore trust as a feeling that emerges when technology is experienced in a broader social and cultural context (Pink et al., 2020) and as a phenomenon that does not solely depend on technological factors.

This thesis takes its starting point in analysing why a move beyond solely focusing on trust in AVs toward working more explicitly with the development of trustworthy intelligent vehicle development can prove to be beneficial for aligning the field of AV development with calls for more reflexive,
contextual and experiential understandings of trust in a growing body of HCAD research. These calls also echo recently developed sociological frameworks for guiding technological development to respond to societal values and needs and consider technology's potentially damaging impact on society and communities (Stilgoe et al., 2013).

To explain this further, this thesis starts with exploring and analysing how research has defined trust and trustworthiness and their underlying logic. In organisational studies, which has influenced trust in automation theory, trust has been defined as “the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party” (Mayer et al., 1995, p. 712). In research on trust in automation, a well-acknowledged definition of trust is “the attitude that an agent will help achieve an individual’s goals in a situation characterised by uncertainty and vulnerability” (Lee & See, 2004, p. 54). These definitions impose that trust is “a psychological event within the individual” (Lewis & Weigert, 1985, p. 967), transactional, and develops in momentary interactions between an individual and technology. Another perspective is that trust is a feeling instead of a cognitive process. Pink et al. (2020) defined trust as “an anticipatory feeling which is sensed when engaging in a practice” (p. 6). They called for trust to develop through people’s experiences with technology in their everyday lives and frequently emerging in uncertain situations (Pink, 2021). Despite how trust is perceived and conceptualised, it is intrapersonal. This means that the responsibility lies within a person.

On the other hand, definitions of trustworthiness shift that responsibility to AV development in a broader social system. Hardin (2006) defined trustworthiness as “a commitment to fulfil another's trust in you” (p. 28), and Hawley (2019) said, “to be trustworthy is to live up to one’s commitments” (p. 23). Meaning to be trustworthy requires intentional effort from the development to understand and account for what mediates people’s trust in AVs. Trustworthiness is also contingent upon stakeholders shaping the vision and conditions for AVs rather than an inherent attribute of the technology. This underscores the importance of a broader framework that brings in the social world in determining trust in AVs instead of sole reliance on the underlying technology. This thesis argues that focusing on trustworthiness complements existing HCI trust research by studying trust in the social and technical systems constituting AVs.

To conclude, two issues form the main argument for why this thesis has chosen to investigate and contribute to the growing body of literature that takes its starting point in human-centred approaches that situate trust in a
broader social context than solely in the direct interaction with technical solutions. First, mainly focusing on technology-related aspects like system performance and design features may inform how to design systems that feel reliable but do not ensure that they align with people’s values and needs and will benefit society. Second, decontextualising automation and AVs from the real world fails to account for the broad spectrum of societal aspects, e.g., laws and regulations, infrastructure, people’s everyday environments, routines, and anticipations, to foresee damaging impacts and prepare society for such technologies. Solely paying attention to how the technology at hand supports interactional trust in the technology does not require to account for a full spectrum of trust influencing factors, e.g., social and cultural aspects that mediate trust when people envision AVs as part of their everyday lives (see, for example, Hoff and Bashir’s (2015) full model of trust influencing factors) which would align the technology with the values and everyday needs of the people affected by it.

Therefore, as argued in this thesis, this shift in focus toward trustworthiness in algorithm design commits to understanding the context AVs will be part of and how trust develops when people envision and experience AVs as part of their daily environments, routines, practices, relationships, and desires.

Aim and Research Question

This thesis aims to contribute by exploring people’s trust in AVs and demonstrating how human-centred approaches can support a more trustworthy and responsible AV development that aligns with societal needs and brings actual value to people. It will do so by providing methodological insights that support shifting AV development from a technology-centred approach towards human-centred practice. This thesis investigates an approach to designing trustworthy AVs by extending the understanding of trust with insights from people’s real-life contexts and demonstrating how human-centric perspectives balance technology-centred views dominating AV development.

With this in mind, this thesis explores “How can human-centred approaches support the development of trustworthy intelligent vehicle technology?” By answering the research question, this thesis contributes to research and practice by:

1. Empirical data that extends the theoretical understanding of trust in AVs with insights from people’s real-life social and cultural context;
Practical insights demonstrating methodological approaches supporting AV development in focusing on trustworthiness, specifically AV algorithm design practice moving beyond the techno-centred approach towards human-centred practice.

The contribution targets researchers and designers in academia and industry. Although this research focuses on AVs, these theoretical and methodological insights can also be transferred to other intelligent technology development contexts.

Thesis Outline
This thesis consists of a cover paper and five individual papers. The cover paper is organised as follows: Chapter 2 introduces the state-of-the-art HCI trust in automation and AVs research, previous research into trustworthiness, and the evolution of human-centred design. Chapter 3 describes the research approach and the development of the research methodology used to study people’s trust in AVs in a real-world context and support AV development, more precisely, the AV algorithm design, to move beyond technosolutionism and become human-centred. In turn, this leads to more trustworthy intelligent technologies. Chapter 4 summarises the five empirical studies that contributed to answering the main research question. Chapter 5 discusses the research approach and the study results and presents the main contribution of the cover paper. In addition, it examines the main limitation and suggests future research. Chapter 6 offers concluding remarks.

Key Concepts
The key concepts in this cover paper that need a brief introduction to how they are used are trust, trustworthiness, autonomous vehicle development, algorithm, trustworthy algorithm design, design ethnography and people.

Trust refers to an intrapersonal trait like the willingness to be vulnerable (Mayer et al., 1995), an attitude that an agent will help (Lee & See, 2004), or a feeling that emerges in a real-life context (Pink, 2021). Trust is one of the central concepts as it is tightly linked to the concept of trustworthiness and, therefore, critical for autonomous vehicle development to become trustworthy. Trust is a concept that is used throughout this thesis.
Trustworthiness is a process, not a product. It demands that different disciplines and sectors continuously and collaboratively learn and evolve. Trustworthiness is the key to shifting AV development from a technology-centred focus to a human-centred practice. To be trustworthy, it is crucial to understand what develops people’s trust. Trustworthiness is a key concept in this thesis.

Autonomous vehicle development refers to the context where citizens, city, industry, and academic stakeholders explore, discuss and design future intelligent transport mobility concepts. All the studies in this thesis focus on different aspects and parts of autonomous vehicle development, either how trust is being conceptualised or how insights from people’s lives are accounted for. It is the leading study context for this thesis.

Algorithm refers to two things. First, guided by Seaver’s (2017) “algorithm as culture” theory, algorithm refers to the enactment of various developers, e.g., people writing requirements for algorithms, validate algorithms, utilise algorithms to develop solutions, the developers who directly or indirectly impact the final algorithm design and behaviour. Algorithms result from activities in many different situations and contexts drawn into action and will have different meanings and roles and generate different practices in different situations. Second, in technical terms, algorithm implies mainly artificial intelligence algorithms but, in some cases, non-AI algorithms. This is because both types of algorithms are used in AVs, and therefore, some studies involved both AI and non-AI algorithm designers.

Trustworthy algorithm design offers four elements - responsive, interventional, intentional and transdisciplinary - that help shift AV development from technology-centred to human-centred and trustworthy practice. These elements are the result of analysing this thesis’s study findings. Trustworthy algorithm design is the main contribution of this thesis.

Human-centred design implies that designers consider the entire experience people have with technology. It calls for reaching beyond experiments that aim at understanding how users use technology to perform tasks to approaches designed to explore people’s routines, values, needs, and anticipations of technologies in real-life situations and environments (Krippendorff, 2004). Human-centred design is central in this thesis as it demonstrates how AV development can become more trustworthy.
Design ethnography is the methodological approach to exploring people’s everyday life context and a “starting point for opening up new perspectives and thinking about new methods that lead in iterative steps to the creation of form” (Müller, 2021, p. 3). This thesis assumes it is an interventional and collaborative research practice (Pink et al., 2022). Design ethnography is the underlying methodological approach for this thesis.

People refer to the members of the public, users, customers, citizens, and residents. In this thesis context, people do not refer to project partners like industry, academic and city stakeholders, or the algorithm designers and AV developers who participated in the different studies.
Related Theory and Research

This chapter first introduces human-centred design and presents related research to bring human-centred approaches to AV development. Secondly, it introduces previous research on trust in automation and AVs. It explores trustworthiness as a concept, its implications for AV development, and why it has proved to help shift the focus from technology-centred understandings of how people accept and adopt technologies toward thinking about trust as something that develops when the technologies become part of people’s everyday lives. Finally, the chapter concludes how this thesis contributes to the research on trustworthy algorithm design that emerges in the intersection of trust, autonomous vehicles and human-centred approaches.

Human-Centred Design of Autonomous Vehicles

This thesis investigates how human-centred approaches can support trustworthy intelligent vehicle development. As will be described in more detail in the coming sections, this implies methods and techniques that decentralise technologies into being part of a broader social and societal context. This approach puts this thesis into a design practice trajectory that goes back at least forty years. Ever since Norman and Draper (1986) coined the term User-Centred Design, there has been a growing understanding of how designing effective technologies requires understanding users' real-world contexts (Giacomin, 2014; Nielsen, 1994; Maguire, 2001). This requires designers to engage with users in their natural environments and observe their interactions with current devices to inform product enhancements (Holtzlag & Beyer, 2014; Karat, 1997; Saffer, 2010). However, the focus on people as ‘users’ has been criticised for reducing human beings to ‘merely cognitive and physical components of a system comprising user, system and environment’” (Jordan, 1999, p. 208). Redström (2006) criticised this way of developing technology as it does not understand what needs to be designed and focuses solely on optimising user task performance based on predictions. When the term Human-Centred Design (HCD) was introduced, it addressed this criticism, foregrounded human needs, capabilities and behaviour, and
suggested designing with these in mind (Norman, 2013). It called to investigate technology use and people’s experiences with technologies in a broader context (Jordan, 1999). This also meant more active engagement with potential users (Maguire, 2001) and consideration of people’s practices, needs and desires (Steen, 2011) in the development processes. HCD suggested that designers consider the entire experience people have with technology. This shift also affected the epistemological underpinnings of the design practice. It called for reaching beyond experiments that aim at understanding how users use technology to perform tasks to approaches designed to explore people’s routines, values, needs, and anticipations of technologies in real-life situations and environments (Krippendorff, 2004). A similar notion was made by Dourish (2006), who suggested that the HCI field adopt approaches illuminating the relationships between technology and people’s everyday practice. More specifically, Dourish proposed to investigate how people put technology into practice in their daily social and cultural context and what meaning technology has in that context. He pointed out that users collectively create the circumstances and contexts where technology is being used instead of passively receiving already defined technologies.

According to HCD principles, to successfully design intelligent vehicle technologies, it is essential to understand the contexts in which AVs are used from within the cultural practices, norms, and meanings of these practices (Dourish, 2007). Designers need to acknowledge and consider these technologies’ impact on people’s everyday real-life context and society and focus on long-term benefits instead of short-term gains (Norman, 2023). This also calls for designing with communities of people (Calvo & De Rosa, 2017) and not only catering for individuals’ needs and values. The issue is that in technology development, often, the goal is to get people to trust and accept intelligent technologies (e.g. Vorm & Combs, 2022) and AVs (e.g. Choi & Ji, 2015) through models that aim at designing the technologies to be perceived useful and easy to use. However, these research investments have been criticised through human-centred approaches for the simplified assumption that the same model can explain people’s behaviour and decisions in different contexts of technology development and use (Bagozzi, 2007). More precisely, what has, among others, been criticised by Bagozzi (2007) and Benbasat and Barki (2007) is that research in this vein neglects methods and theories that explain why people perceive technology to be valuable and easy to use; it fails to account for social and cultural aspects that guide people’s decisions to use technology, and it oversimplifies understandings of people’s feelings and emotions. This criticism aligns with HCD calling for technology development to change its focus towards understand-
ing the social and cultural context in which technology will partake and its impact on those contexts.

This is ever more prevalent nowadays when algorithms are increasingly integrated into society while often causing harm in real-world applications at the expense of people (Kak & West, 2023) and coming up with incomprehensible results even to their designers (Gunning & Aha, 2019). Focusing solely on technology and user acceptance does not ensure long-term relationships with people. Therefore, it is not enough in the HCD of AVs to focus exclusively on manipulating technological factors that develop people’s trust in technology without considering the impact the technology can have on their lives. In fact, HCD methodologies can support AV development to become more intentional in understanding the social and cultural context, as Klaus Krippendorff explains:

“human-centredness takes seriously the premise that human understanding and behaviour goes hand-in-glove; that what artefacts are is inseparably linked to how their users perceive them, can imagine interfacing with them, use them and talk about their stake in them with others” (Krippendorff, 2004, p. 48).

The development of human-centred design has spawned several different approaches that have inspired the work of this thesis. In the next section, these approaches will be described to how they have been integrated into human-centred AV design.

Human-Centred Autonomous Vehicle Design Approaches

A growing body of literature relevant to this thesis reports on human-centred approaches concerned with interpreting and explaining people’s experiences with autonomous vehicles in real-life contexts or creating theories from empirical data on people’s doings and whereabouts in everyday logistics. For example, *ethnomethodology* aims to study and learn about people’s naturally occurring ordinary activities (Garfinkel, 1967; Crabtree et al., 2012). It seeks to uncover people’s actions to make sense of and structure their social lives (Button et al., 2015). Ethnomethodology has also been applied to investigate AVs in real-world environments. Brown et al. (2023) analysed AV videos to study overlooked aspects of driving interactions. They analysed publicly available videos of AVs and focused on a rudimentary element of road interaction: how people, when they drive, give way to each other. They discovered that AVs still struggle with such elementary actions. They concluded that understanding traffic has an essential role in
AV development, and therefore, a significant challenge is not technological but understanding social interactions. Pelikan (2021) investigated how autonomous shuttle buses coordinate with other road users, like pedestrians and cyclists, in unstructured everyday spaces. She conducted her research on a university campus where the buses were tested. She found that the buses failed to coordinate their movement in regular traffic. She argued that if AVs are expected to cope with regular traffic, they must also be designed to handle mundane interactions with humans, not only smooth traffic scenarios. She emphasised the need to study AVs in real-world traffic and everyday situations.

Other relevant human-centred approaches have been developed in response to criticisms that design for societal needs sometimes excludes the most relevant stakeholders (the users) from the design process (Binder et al., 2008) and, therefore, focus on involving users and other local stakeholders in the design process. Many of these design-oriented approaches are described under the umbrella term design thinking, presented by Brown (2008) as a “discipline that uses the designer’s sensibility and methods to match people’s needs with what is technologically feasible and what a viable business strategy can convert into customer value and market opportunity” (p. 2). It consists of understanding user needs, creating solutions for them, and evaluating these solutions with users (Foster, 2021). Design thinking originates from design practice but has also gained wide popularity in technology development (Cross, 2023). Sirkin et al. (2016) used design thinking to uncover implications for designing automated vehicle interactions with in-car passengers and other road users. They engaged participants in three phases: conceptualisation, prototyping and fieldwork with a deployed AV prototype. They demonstrated how design thinking can be used to investigate people’s reactions to technologies that do not exist yet. Oehl et al. (2020) report on an expert workshop with researchers and practitioners working with affective interfaces for AVs. The participants were engaged in design thinking to develop use cases for different AV use scenarios to offer direction for implementing affective technologies into AVs. They concluded that user needs must be collected from real users and might differ based on the cultural contexts. Manoeva et al. (2022) employed design thinking to examine, from the user’s perspective, how autonomous three-wheeled cargo bike on-demand sharing service should be designed to satisfy people’s expectations and needs. They argued it to be a valid way to involve users in different parts of an innovation process to increase the probability of people accepting a solution.

Co-design is a human-centred approach that actively engages stakeholders to design solutions for pre-existing problems (Messiha et al., 2023).
Porter and Porter (2000) stated that “co-designing is essentially design with, for and about users” (p. 129). Co-design allows the combination of the knowledge from the users with the knowledge of professionals, which can lead to novel ideas (Trischler et al., 2018). For instance, Asha et al. (2021) utilised a co-design approach to explore and prototype interfaces that support wheelchair users to cross a street in the context of AVs. Over five weeks, they engaged with a powered wheelchair user to understand their lived experiences of crossing streets. Together with the participants, they developed several interface prototypes to be installed on AVs, street infrastructure and wheelchairs to support interactions between AVs and wheelchair users. They revealed that people in wheelchairs might need unique interfaces that help them share streets with AVs but do not want solutions that draw them extra attention. Arvola et al. (2023) also engaged with differently abled participants. Their study focused on the participants’ experiences in autonomous shuttle buses. They engaged five participants who had experienced riding an autonomous shuttle bus in co-design workshops to explore how the participants would interact with such a bus in different situations. Their study resulted in implications for an explainable autonomous shuttle-bus interior that accounts for temporal aspects of passengers’ behaviours.

Participatory design is a human-centred methodology that actively engages people in new technology design and decision processes (Muller & Kuhn, 1993). It is led by two values: people have a democratic right to participate in the design process, and there is an opportunity for design to harvest “tacit knowledge” from participants and their formal competence (Ehn, 2008). Participatory design is widely applied in intelligent vehicle technology and transport mobility research. For instance, Riggs et al. (2019) assembled stakeholders from automotive engineers, urban planners and policymakers to explore how to design streets for AVs and promote urban living. A common theme that emerged from the study was that future streets need to shift from being car-centric to people-centric, and a large part of them need to accommodate other road users like pedestrians and cyclists while being easily understandable and navigable for the vehicles. Brinkley et al. (2022) argued that there is a lack of knowledge on designing for accessibility interaction in AVs. To fill that gap, they utilised participatory design to engage elderly adults and people with visual impairment in the interaction design for AVs. They elicited 12 accessibility guidelines for inclusive technology development and demonstrated the value of participatory design in extending AV design from commonly involved users to include people with different characteristics. Carvalho et al. (2021) utilised participatory design to involve elderly adults and people with special needs as co-designers for
AV ride-sharing applications. They discovered that the needs and requirements of the participants mainly revolved around issues of trust, safety, reliability and ease of use of AV interfaces. They concluded that AV development needs to consider accessibility as a crucial part of the design process rather than an afterthought. Al-Taie et al. (2023) engaged cyclists in designing external AV interfaces for interactions with cyclists. They argued that previous research on external AV interfaces has mainly focused on vehicle and pedestrian interactions and lacks insight into AV and cyclist interactions. From the cyclist’s preferences to external AV interfaces, they developed a taxonomy that led to four AV interaction design concepts.

As the examples above demonstrate, human-centred approaches have inspired many techniques and methods to emerge in the collaborative, participatory and speculative design of AVs and other intelligent technologies. For this thesis and the nature of the externally funded projects the studies presented are connected to, these approaches were combined and framed by a methodology that has proven valuable in informing design and staging deeper engagements with people and places, namely design ethnography (DE). DE is a human-centred approach that seeks understanding people’s routines, methods and anticipations in real-life contexts to inform design. In this process, the ethnographic encounters become embedded into an iterative design process. Through long-term engagements with particular social contexts and communities, DE offers analytical accounts of an in-depth understanding of peoples’ experiences, meaning-making and anticipations of technology in their everyday social and cultural context that stretches beyond more conventional design iterations to check task performance and usability in everyday contexts (Müller, 2021). For example, Pink et al. (2017) conducted in-car ethnography to understand aspects of car-based mobility that could be used to inform AV development. They concluded that AV development should consider driving as an interplay between drivers, vehicles and the environment, not solely human-machine interaction. In addition, the results demonstrated how people’s past and present experiences are intertwined with how people envision and anticipate future experiences with AVs. Another example is Latham and Nattrass’s (2019) ethnographic exploration of how people share and cooperate within a street and the challenges in AV development to make them cope with everyday urban streets. They reached an understanding that people cooperate and coordinate streets depending on a mix of aspects like physical infrastructure, regulations, people they share roads with and rules of conduct to conclude that AV development needs to be sensitive to the infrastructural arrangements that define different environments and contexts. Eden et al. (2017) conducted an ethnographic study with stakeholders like shopkeepers, drivers, pedestrians
and residents to understand their perceptions and attitudes toward AVs. They focused on how people interact and make sense of AVs in public spaces. They argued for how ethnographic research is equipped to calibrate AV technology to account for how people coordinate themselves with others when sharing roads. Lindgren et al. (2022) adopted a design ethnographic approach to explore peoples’ changing anticipations towards AVs. For 18 months, they engaged with families using automated driver assistant functions. They also immersed the participants in AV scenarios using Wizard-of-Oz prototype cars. Through these techniques, they could learn how people adopt and anticipate experiencing non-existing technologies like AVs based on everyday life practices and circumstances. DE has also been proven to be used to identify and reframe taken-for-granted assumptions in future mobility debates and discourse. For instance, Brodersen et al. (2023) undertook a DE study to challenge a commonly adopted attitude that the first and last mile of peoples’ everyday logistics is primarily grounded in time and cost efficiency that can be enforced through different service solutions that make the travels quicker and more straight forward. This idea could be contested through a participatory DE approach that explained how the first and last mile of people’s travels, on the contrary, was filled with social endeavours far from grand narratives of seamless and efficient multi-modal commuting. This generated an understanding of mobility’s collective and social nature from people’s everyday lives and suggested future mobility solutions to account for the nuances and contingencies that people must deal with in everyday mobilities.

In the above sections, human-centred design and approaches to developing intelligent vehicles have been discussed for this thesis. The following section presents and analyses research on trust and arguments for how human-centred approaches are equipped to nudge this field toward what is here called trustworthy algorithm design.

Trust in Autonomous Vehicles

Trust is one of the key factors influencing people’s decision to use or not to use automation (Muir, 1987; Parasuraman & Riley, 1997) and is, therefore, essential to the success of AVs (Wintersberger & Riener, 2016). This is mainly because AVs are anticipated to bring societal benefits like increasing traffic safety, providing mobility for non-drivers, and decreasing the environmental impact of vehicles (Litman, 2021), which would require people to use AVs. HCI is a field that has shown increased interest in researching trust between humans and automation, e.g., AVs. There has been an exponential
growth in published trust in automation and AVs research articles since the early 2000s. For example, articles that report on empirical studies on trust in command-and-control systems (see, for instance, Sheridan, 1988), human-robot interactions (see, for example, Kaniarasu & Steinfield, 2014), and trust in AVs (see, for instance, Ekman, 2023; Raats et al., 2020).

Previous HCI research on trust in automation and AVs have contributed with several models (see, for example, Muir, 1994; Lee & See, 2004; Choi & Ji, 2015; Hoff & Bashir, 2015), frameworks (see, for example, Ekman et al., 2018; Moran et al., 2019), questionnaires (see for example, Lee and Moray, 1992; Muir & Moray, 1996; Jian et al., 1998), and definitions (see for example, Muir, 1994; Lee & See, 2004) that help to explain, understand and study trust in automation. To give an overview of these contributions, this section describes how trust has been understood and defined on a general level. Then, it explains how the different understandings have been modelled and how the models have evolved in recent history. It then details empirical studies on trust in automation and AVs and discusses the study setups, trust assessment, and aspects of trust that have been in focus. Finally, it opens up broadening the understanding of trust in automation.

A common understanding among HCI trust in automation researchers is that trust is always placed in someone or something (Muir, 1994). Trust is based on the trustee’s information on the trustee and, therefore, can differ between trustees, e.g., one may trust a recommender system but not an AV. As it is not always reasonable or even possible to fully understand how a system works, it is believed that trust helps to cope with complexity and uncertainty (Lee & See, 2004). For example, it is impractical for people to understand how AVs work entirely; even the developers designing (AV) algorithms cannot always explain why an algorithm comes up with specific results. This speaks to Pink et al. (2018), who considered trust a feeling that allows people to move on in situations where absolute certainty is impossible. This is the case with intelligent technologies, e.g., AVs, which are not fully comprehensible due to the nature of algorithms. Trust also depends on the context in which people interact with a system (Lee & See, 2004); e.g., a driver-assist system that helps to keep a vehicle in a lane can be trusted on the road with clear lane markings compared to a dirt road. Finally, as trust mediates reliability, it tends to be more critical in situations of uncertainty and vulnerability (Muir, 1994). One of the most commonly accepted definitions in HCI trust research has been that trust is “the attitude that an agent will help achieve an individual’s goals in a situation characterised by uncertainty and vulnerability” (Lee & See, 2004, p. 54). Although trust does not pose a risk, it is the willingness to take risks, which means one can trust an AV but never ride in one.
In 1994, Muir pointed out a lack of theoretical and empirical work on trust in automation and that measuring trust in any direct sense is impossible. She developed the first trust-in-automation model to fill this gap and support empirical studies on trust in automation. The model attempted to bring notions of interpersonal trust into trust in automation research, clarify trust terminology, and discriminate between concepts often used together in trust research, e.g., trust, predictability and accuracy. Muir’s model described the relationship between automation and a user’s trust. The model showed a connection between the user’s mental model and the actual characteristics and behaviour of automation and a continuous loop between the user’s predictions about an automation behaviour and its effect on trust. A decade later, Lee and See (2004) offered an extended model by emphasising the importance of the context in such relationships while elaborating on the impact of the user’s experience with a system and appropriate trust calibration. The subsequent development came with Hoff and Bashir’s (2015) full model of factors influencing trust in automation. This was significant as while further detailing the role of context in trust development, they also proposed trust to have three layers - dispositional, situational and learned - extending the understanding of trust in automation further.

Overall, the fundamental understanding of trust in automation proposed by these theories is that trust is transactional, developing through momentary interactions between a user and a machine. These models (see Figures 1, 2 and 3) have had an essential effect on trust in automation and AVs studies. They created a foundation for conceptualising and conceptually studying trust in automation and AVs. A further explanation of this will be given in the following discussion about a systematic literature review on HCI trust research, also reported in Raats et al. (2020) (Paper 1). Although first, it is important to mention another significant contribution, namely Muir and Moray’s (1996) empirical testing of Muir’s (1994) trust model. They concluded competence to be a strong indicator of users’ trust in automation. They said, “While it is useful to know that competence is an important determinant of trust in a machine, it is necessary to examine the expectation of competence more closely if we are to develop a predictive model of trust in machines and intervention behaviour” (p. 456). Competence is the ability of a system to perform as expected in a given situation and context. Looking back, this finding could be attributed to why other HCI trust in automation and AVs research mainly focuses on system performance factors and their impact on trust development.
Figure 1. Model of the relationship between the automation, the operator's trust, and predictions about the automation's behavior (Muir, 1994).

Figure 2. A conceptual model of the dynamic process that governs trust and its effect on reliance (Lee & See, 2004).
In Raats et al. (2020), the analysis of 258 reported studies on trust in automated vehicles revealed three critical areas of understanding that, by the time the review was conducted, seemed to have been neglected in research: a) to understand if learnings of trust generated in simulated scenarios also apply in real-world contexts, b) to study trust in real-life circumstances to extend previous knowledge of trust, and finally, c) to explore trust through interdisciplinary and human-centred, participatory and ethnographic approaches that are designed for accounting a more comprehensive network of trust relations than merely trust between a person in an AV. More precisely, the dominant body of studies in the review was conducted in laboratory setups, test tracks and computer simulations (209 out of 258 studies) with participants who were not considered the primary end users of the tested systems. Only six studies had been conducted with people in their everyday situations and five on public roads in real traffic. Trust had been commonly assessed through questionnaires (221 studies out of 258), e.g., the Trust in Automated Systems Scale (Jian et al., 2000), in contrast to only one study utilising ethnography to explore trust in peoples’ real-life environments and circumstances. The research predominantly focused on how technology-centric aspects like design features (104 studies), e.g., appearance, ease-of-use and system feedback and system performance (91 studies), e.g., reliability, predictability and usefulness of a system influence trust development. Despite the research offering a significant amount of knowledge on how
trust develops when people interact with automation and AVs, the empirical studies neglected to explore the impact of the social and cultural context where these interactions would partake in trust development. These findings resonate with the theory of trust in AVs, and the conclusions made by Muir and Moray (1996) on competence, also known as system performance, is a strong indicator of user trust. This also shows the opportunities for extending trust theory by understanding trust from real-life contexts and new interdisciplinary approaches for studying trust in AVs. Figure 4 is an illustration from Raats et al. (2020), which uses Hoff and Bashir’s (2015) full trust in automation model to explain what aspects of the model usually is targeted by HCI trust research and which elements in the model could be captured with developed ethnographic methods.

In the same vein as Hoff and Bashir (2015), Muir (1994) developed a trust model that shows how trust develops in an individual’s interactions with automation to support empirically studying trust while at the same time proposing that “trust is only part of a network of trust that pervades complex, automated systems.” (p. 1906) (see Figure 5 for Muir’s (1994) trust network illustration). Also, her definition of “trust is the expectation, held by a member of a system, of persistence of the natural and moral social orders, and of technically competent performance, and of fiduciary responsibility, from a member of the system, and is related to, but is not necessarily isomorphic with, objective measures of these properties” (p. 1911) indicated trust to be a complex multidimensional concept. For example, it as-

Figure 4. Illustration of how ethnographic approaches could be helpful in studying factors of dispositional and situational trust not usually addressed by conventional HCI trust research approaches. The illustration is based on Hoff and Bashir’s (2015) full trust in automation model from Raats et al., (2020, p. 9).
sumed trust as part of a system and considered trustees to have a moral commitment to act decently. The latter spoke to interpersonal trust theory, which defined a person’s trustworthiness as a commitment to fulfil another’s trust in them (Hardin, 2002; Hawley, 2019). Perhaps these notions have not been picked up by trust in automation studies because industrial automation has a concrete task to perform, which can be predicted and validated. Whereas automation, like AVs, anticipated to increase traffic safety, provide mobility for non-drivers, and decrease environmental footprint (Litman, 2021), would benefit from this way of thinking, as AVs are expected to operate in uncontrolled environments and real-life contingent situations, therefore posing more risk. This resonates with Muir’s (1994) notion of trust being part of a network. The underlying AV technology does not carry the commitment to fulfil trust, as inanimate objects cannot have moral obligations to act reasonably, but by the assembly of different actors who develop these vehicles, e.g., developers, designers, city infrastructure architects, and lawmakers. In addition, it acknowledges the calls for considering trust as a feeling experienced in everyday life contexts (Pink et al., 2020) rather than solely something to be invested in during an interaction.

![Network of trust](Muir, 1994)
Defining trustworthiness

Based on the overview of the development of models of trust in automation over the past two decades, this thesis argues that human-centred approaches to AV development could benefit the field by focusing on designing trustworthy technologies beyond solely inter/transactional trust in order to contribute to the development of technologies that are not only accepted by people but also align with people's values and needs thus bringing wider societal benefits into the equation. Trustworthiness is a concept that has yet to receive consideration in HCI research on interactional trust in automated vehicles. On the other hand, trustworthiness has been studied more in detail in relation to interpersonal trust research. Hardin (2006) defined trustworthiness as “a commitment to fulfil another's trust in you” (Hardin, 2006, p. 28), and Hawley (2019) stated that “to be trustworthy is to live up to one’s commitments” (p. 23), it is an aspiration. This thesis adopts and tailors this definition of trustworthiness to demonstrate how human-centred approaches can support AV development in becoming more trustworthy.

The problem is that trustworthiness does not apply to inanimate objects as they cannot have moral commitments or interests to act reasonably towards people. This is why it is essential to consider AVs not solely as intelligent technologies but as transdisciplinary endeavours. Transdisciplinarity means that relevant expertise and perspectives like the public, government, academia, and industry assemble to solve real-life problems (Polk, 2015), and by doing that, collectively learn and enhance each other (Jantsch, 1972). AVs will impact how people move around and organise their lives, will require new laws and regulations that will take care of people in case the vehicles are in accidents, change how cities are planned, and, in the worst case, amplify societal issues like inequality in access to public transportation. During the last year, an influential strand of social sciences research has emerged that claims the relevance and importance of implementing frameworks that support cross-sectoral and interdisciplinary reflective practises to make intelligent technology development and innovation in responsible ways (Roy, 2021). This means reflecting on the assumptions and attitudes implemented into future solutions, anticipating a solution's impact on society with a heterogeneous set of stakeholders (including affected citizens) and responding to societal values and needs (Stilgoe et al., 2013). This aligns with the notions of trustworthy technology development outlined in this thesis. It contrasts and complements more technology-driven logic in refining technologies to make them easier to use and understand (Davis, 1989). This shift in mindset (from making people trust and adopt technology to support the development of trustworthy technologies) concerns nudging
different stakeholders into collaborating to create societal value with the help of technology rather than trying to get people to accept ready-made technologies that are to be implemented. Subsequently, this means that the different stakeholders partaking in trustworthy AV development benefit from learning from each other and learning and changing their ways of development. This implies that people’s trust in AVs needs to be explored not only towards the underlying AV technology but also a more extensive trust network of, e.g., developers, designers, city infrastructure architects, and lawmakers (Muir, 1994) and as it develops when AVs partake in people’s real-life social and cultural circumstances and environments. Thus, the aspiration and commitment to be trustworthy are not held by AV technology but by the stakeholders that create the visions and conditions for AVs. This implies that people’s trust depends on the networks of people, places and things that AVs become part of, not solely on, e.g., the underlying AV technology. O’Neill (2018) made a similar implication by suggesting investigating the influence of expertise, professionalism and evidence on trustworthiness. This indicates that trust is a multilevel concept of values, principles, practices, stakeholders, and artefacts.

Focusing on trustworthiness complements HCI trust research through technology-related aspects by adding ambitions to study trust in the systems of people, places, processes and things suggested by Muir (1994) that constitute the development of AVs. Simon and Rieder (2021) made a similar conclusion in their analysis of a Corona-Warn-App’s trustworthiness. They hinted at people’s trust being placed not only on the app itself but also in the people developing the app, their values and the principles that guide the development. Also, Shneiderman (2020) claimed that trustworthy AI development would need practices supporting technology reliability, organizational strategies that nurture culture and safety, and self-governed oversight supporting trust. Focusing only on trust runs the risk of people developing trust in AVs under a false pretence (Hawley, 2019) based on the information crafted to do exactly that, to induce trust. However, it would not necessarily make AVs trustworthy, as it has been demonstrated through algorithms causing harm in real-world applications. In contrast, aiming at being trustworthy requires exploring trust as part of a complex network of trust relationships (Muir, 1994) and focusing on individual interactions with technology and, how technology is put into practice and what meaning it has to people (Dourish, 2006).

Furthermore, Hardin (2002) explained that trustworthiness means valuing and wanting to continue relationships with people who trust you. Focusing on trustworthiness would mean AV developers care for the people and communities the AVs will impact. This is essential, as to gain from the an-
ticipated societal benefits of AVs, e.g., reduced traffic accidents, access to mobility for non-drivers, and decreased environmental impact (Litman, 2021), people need to keep trusting the AVs. AVs must be useful and bring actual value to people and their lives. It is only then that the anticipated societal benefits can be gained. Hardin (2002) called this the encapsulated interest. This means the interests of the ones trusting are also part of the interests of the one being trusted. Verberne et al. (2012) made a similar conclusion about a driver-assistance system called adaptive cruise control in the context of vehicle automation. They concluded that drivers found adaptive cruise control more trustworthy due to shared driving goals, e.g., driving comfort, safety, energy efficiency, and speed. The motivation for being trustworthy is maintaining relationships with trusters to fulfil your own interests. Engineering agendas and industry visions of societal benefits mainly drive AV development.

The crux remains to figure out how to implement these understandings of developing trustworthy intelligent technologies. Despite the value of human-centred, participatory, collaborative and responsible technology development being acknowledged on a societal level, few studies have contributed with answers on how to facilitate this reflective technology development in practice, which includes invested stakeholders such as the actual technology designers and developers, the city infrastructure strategists, the affected citizens and research. Reijers (2020) attributes this lack to the dominant focus in innovation studies to control the impact of innovation instead of developing the co-creative activities that lead to it, even though co-creation has been growingly acknowledged as a way to actualise these wanted qualities into technological solutions (Robinson et al., 2021). However, although examples of how to facilitate co-creation activities to ignite trustworthy design of intelligent vehicle technologies are rare in technological trust research, there are some exceptions. For example, technology co-creation has been facilitated through, e.g. card-based tools to trigger collaborative reflections in technology development (Urquhart & Craigon, 2021), photography to engage citizens in urban developments (Altamirano-Allende & Selin, 2016), and co-creation to involve stakeholders in identifying areas where nanotechnology could bring value (Jansma et al., 2022). Thus, co-creation has been demonstrated as a viable way forward to mitigate the risk of AVs being developed following solely engineering visions and agendas that risk causing harm in real-world applications if not the development process has included reflections about societal and social implications of the implementation of intelligent vehicles. Therefore, AV development could be advanced by adopting approaches that support involving people in the development process, understanding people’s everyday life routines, practices
and anticipations, foreseeing possible harmful impacts of technology on society, and responding to the actual societal values and needs.

In summary, trustworthiness is a virtue that AV technology development should aspire for. Trustworthy AVs are human-centred as they are designed with the people and communities they will impact and with consideration of what the impact will be. According to previous research presented above, the design of trustworthy AV technologies calls for two new research directions to be even more developed. First, HCI trust in automation and AVs research needs to acknowledge and research trust not only as a cognitive process happening within an individual (Lewis & Weigert, 1985, p. 967) but also as a frequently emerging feeling (Pink et al., 2020; Pink, 2021) when people envision and experience technology in real-life environments and circumstances. Furthermore, trust needs to be explored as part of a more extensive trust network (Muir, 1994) instead of the user’s individual interactions with technology. Second, the development must acknowledge, account for, and design with the people, communities, and various other societal actors impacted by the AVs to foresee any harmful consequences, solve correct problems, and bring actual societal value. This thesis contribution straddles these two avenues of research by pulling together trust research with multi-stakeholder co-creation of AVs through a human-centred approach outlined in the next chapter.

Towards Understanding Trustworthy Algorithm Design

In conclusion, various human-centred methodologies have been used to inform intelligent technology development for future transport mobility with social insights or, more concretely, design implications and to engage participants in the development process. Previous research has identified a lack of contextual and real-world understanding of people’s attitudes, anticipations and experiences with AVs and a need to expand who is involved in the AV development, e.g., elderly adults, differently-abled people, and cyclists. Researchers have also utilised human-centred research approaches to challenge and reframe commonly adopted assumptions and dominant narratives around the value of automation, efficiency and personalisation of future mobility systems (Cohen et al., 2020).

Therefore, scholars have called for AV development to adopt and adapt more human-centred approaches that support eliciting people-centric and contextual insights from real-world environments and situations to balance technology-centred views and assumption-based decisions that seem to drive AV development. In turn, this helps to increase people’s trust and
acceptance and move towards achieving the anticipated societal benefits (Litman, 2021) that are dominantly used to argue why to develop AVs in the first place. This thesis will answer this call by presenting a deeper understanding of how human-centred approaches can support the development of trustworthy intelligent vehicle technology, focusing on the core of these technologies, namely the AI-driven algorithms that instruct the systems to perform independent operations. However, to capture how algorithms are part of a network of processes, things, places and people, they are seen as more than coded descriptions guiding computational procedures. Instead, they are viewed as a cultural force (Seaver, 2017), shaping traits without a fixed definition. Seaver’s perspective on algorithms as culture posits that they arise from varied situations and contexts, influencing practices. Consequently, algorithms' meanings, roles, and practices diverge based on developers, activities, and applications, potentially leading to contradictions (Dourish, 2006; Seaver, 2017).

This perspective on algorithms resonates with recent initiatives in understanding algorithm design practice. Algorithms reflect their designers’ assumptions, attitudes and beliefs (Allen et al., 2006), and when they cause harm in real-world applications, it is at the expense of people and their lives (Rooksby et al., 2009). However, the designers often cannot explain why an algorithm produces a particular result (Gunning & Aha, 2019). Inspired by human-centred perspectives and to mitigate these risks, Baumer (2017) argues for advancing algorithm design by implementing human-centred strategies in algorithm design practice. Baumer has combined various initiatives (see, e.g., Amershi et al., 2014; Nafus, 2018; Rader & Gray, 2015) to engage users and other stakeholders in the algorithm design process under the umbrella he called the human-centred algorithm design (HCAD) framework. HCAD suggests algorithm designers apply theoretical, participatory and speculative strategies to align their intentions with people’s perceptions of the meaning of algorithms and limit algorithms causing damaging societal impact. Baumer’s framework has inspired further algorithm design (e.g., Cherrington et al., 2020; Dekker et al., 2022; Flügge et al., 2020) to apply theoretical strategies to inform algorithm design with a deep understanding of people’s everyday practices, routines, relationships, and anticipations. Furthermore, Baumer suggests using participatory approaches to stage collaborations between heterogeneous sets of stakeholders to ensure that algorithms solve correct problems and adopt speculative techniques to foresee algorithm failures and unexpected societal impacts. These strategies resonate with calls for focused attention, understanding and consideration of people and society in technology development.
Through the lens provided by Seaver and Baumer and the analysis of previous work presented above, this thesis focuses on understanding the intersection of research on trust, autonomous vehicle development and human-centred approaches as a novel and timely field in need of further investigations: *trustworthy algorithm design* (see Figure 6). By doing this, the ambition is to add to a deeper understanding of how human-centred approaches can actively participate in the development of design methodologies that are attuned and tailored to facilitate the emerging turn toward responsible, accountable and socially relevant AI-driven technologies.

![Diagram](image)

*Figure 6. This thesis aims at developing Trustworthy Algorithm Design by bringing together research on trust, autonomous vehicle development with human-centred approaches to technology design.*

The next chapter will demonstrate how a DE-inspired methodology has developed during the life of this thesis project, tailored to explore the intersections between trust, AV development, and human-centred approaches.
Research Methodology

The following section introduces DE as the underlying research methodology utilised for this thesis, how the methodology evolved through the five studies, and the techniques used to answer the overarching research question, “How can human-centred approaches support the development of trustworthy intelligent vehicle technology?”. The section will continue summarising the five appended research papers and finish by discussing the research quality and ethics.

Design Ethnography

As mentioned in the previous chapter, Design ethnography is a fusion of ethnography and design disciplines to enhance and acknowledge understanding of the social and cultural context being engaged with (Müller, 2021). To produce that knowledge, design ethnographers use ethnographic techniques to participate with the people involved or affected by the design within their context (Elliott & Culhane, 2017). However, ethnography is also shaped by the discipline it is engaged through (Pink & Morgan, 2013). For instance, Baskerville and Myers (2015) establish design ethnography in Information Systems research fields by distinguishing between three ways of utilising ethnography in design. First, to inform design by generating a better understanding of the people to be using the design outcomes. This way assumes ethnography as a technique to improve designs. Second, to study the designers and the design process itself. This aims to better understand the circumstances of the designers as essential participants in the design practice. Third, what they call DE is descriptive and generative. It is concerned with understanding people and their context to make better designs and engages ethnographers actively in design activities.

On the other hand, Pink (2022) establish DE within anthropology by combining the orientation towards possible futures through design with anthropologically theorising the everyday experience, imaginaries, and contingencies as playing out in an ongoing emergent world. Thus, DE should not be considered an enclosed method but a base for developing new meth-
ods in relation to disciplines and contexts that can lead to more socially relevant designs (Müller, 2021). Pink et al. (2022) capture this by describing DE as a “methodologically adventurous, rapidly adaptive and theoretically robust approach to interventional research practice and engagement.” (p. 2). Therefore, DE aims not only to produce conventional design outcomes. Instead, it places interdisciplinary collaboration at its core and stages interventions to imagine unknown futures (Pink et al., 2022). Practically, ethnographic research assumes long-term immersion in other people’s lives (Müller, 2021). However, in more applied contexts like design research, these encounters might as well be shorter and more intensive (Pink & Morgan, 2013). They utilise observational and interventional techniques to elicit what matters to the people in the explored context.

This thesis is inspired by Pink et al.’s (2022) notion of DE as an interdisciplinary practice that seeks to integrate ethnographic research with design practice. This can only indeed be done by taking part in situations where futures are being imagined and shaped by different disciplines and stakeholders. Only then is it possible to critique, challenge and shift ideas and participate in defining and co-creating alternative possible futures (ibid.). These engagements also create a site for collaborative meaning-making and knowledge production (Pink & Morgan, 2013), with a starting point in ethnographic understanding. Altogether, Pink (2009) conceptualises this mingling of ethnographic materials, analysis and multi-stakeholder meaning-making as the ethnographic place that requires ethnographic data to be tailored to represent the research participants, address the audience, and facilitate knowledge sharing and collaborative learning. These are the places where ethnographic data become instruments of intervention.

As this thesis takes a stance of DE being an interventional and collaborative approach, the research outcomes are not bound to this thesis nor the overarching projects. Instead, they intend to have a broader impact on society and technology innovation that cannot be imagined or anticipated (Pink et al., 2022). Such engagements are undertaken through activities planned within the overarching projects but facing outwards. For instance, engaging outside participants through workshop facilitations at international conferences, panel discussions at network events, research share-outs at fellow academic institutions or industry presentations.

Data Production

This thesis data was produced over four years within two externally funded collaborative projects, Trust in Intelligent Cars (TIC) and Design Ethno-
graphic Living Labs for Future Urban Mobility – A Human Approach (AHA) (see Figure 7).

The TIC project was an interdisciplinary project oriented towards the automotive industry. Two studies reported in this thesis were conducted within this project. The first study (reported by Raats et al. (2020) (Paper 1)) investigated how the increased vehicle intelligence affects people’s and public trust in AVs. It answered the question based on previous HCI literature on state-of-the-art HCI trust in automation and AVs research while revealing a research gap and limited understanding of what mediates trust when AVs are envisioned in people’s real-life context. Guided by the notion that DE enables collaborative engagements (Pink et al., 2022), the second study (reported in Raats et al., (Submitted, 2023) (Paper 3)) undertook what Baskerville and Myers (2015) call ‘ethnography to study design’ and investigated where to stage such engagements in AV development. Through the work in the second study, a focus on AV algorithm design was established as it became clear through the results that while AVs rely on the advancements in AI technology (Nascimento et al., 2020), the behaviour of the algorithms that the designers’ attitudes and assumptions impact power AI (Allen et al.,

---

**Figure 7. Overview of the research projects, studies and design ethnographic techniques used in data production.**
However, it became evident that there is limited knowledge of how AV algorithm design is practised. The second study, therefore, concentrated on learning how human-centred approaches and societal needs and values accounted for in AV algorithm design. This study was carried out through ethnographic observations and interviews at two offices of an automotive company, including participating in different work and social activities. Spending time with the participants, although not actively engaging in the discussions or development of AVs, created an understanding of the dynamics and conditions that shape algorithm design practice for AVs. This knowledge was crucial for later studies in the AHA project, where further engagements with AV algorithm designers were tailored and staged. The observations and interviews were documented through descriptive field notes and transcribed audio recordings.

The AHA project was an interdisciplinary project structured by DE methodology (Pink et al., 2022). It focused on interventional and future-looking engagements with citizens, city, industry and academic stakeholders in the context of future urban mobility. The project concentrated on two semi-rural areas in Sweden, Drottninghög, Helsingborg and Bergum-Gunnilse, Gothenburg. Drottninghög is a closed-off area of identical apartment complexes. The population is mainly with migratory backgrounds. The area is connected to the closest city centre, Helsingborg, via public transport. Bergum-Gunnilse is composed of clusters of residential housing. It is a widely spread-out area on a hilly landscape. The clusters are connected by a highway connecting the areas to the closest city, Gothenburg. The primary way to commute to the city is via individually owned cars or public transportation that can be accessed on the highway. One important element of the AHA project was establishing DE Urban Living Labs in these areas, comprised of a network of supporting activities that engages local stakeholders, citizens and project collaborators in exploring future mobilities. It involved engaging in the areas over three years using ethnographic methods to share citizens’ experiences of moving around in the areas and their imagined mobility futures. It also included working with stakeholders in workshops to co-create knowledge, imagine future technologies and co-design prototypes and services. Through these various research and other project activities, the AHA project offered a stage to explore and experiment with different DE techniques according to DE principles. That is, tailoring techniques to uncover what matters to people in these particular areas and how to engage participants in making sense of the DE materials, using it to imagine alternative AV futures and turning them into instruments for future explorations. For instance, study three (reported in Raats et al., (Submitted, 2022) (Paper 2)) used probes (Mattelmäki, 2008) to explore elements of
trust when people envision AVs as part of their everyday mobility alternatives. The study participants were people already sharing and organising their daily lives with each other in the Bergum-Gunnilse area. The study was conducted through online workshops to comply with COVID-19 pandemic restrictions, and the workshop agenda was based on results from a pre-interview done with the workshop participants. Digital maps of the participants’ areas and visualisations of AVs were used to facilitate participant engagement and situate AVs in their concrete environments. This supported the participants in envisioning and explaining their area in relation to AVs. Online collaboration platform Mural facilitated the collaborative engagement with the digital maps. The fourth (reported in Raats et al., (2023) (Paper 4)) and fifth studies (reported in Raats (2024) (Paper 5)) were guided by the notion that DE is an interventional practice holding interdisciplinary collaborations at its heart (Pink et al., 2022). These studies focused on creating an evolving ethnographic place (Pink, 2009) for participants to participate in when collaboratively creating possible AV futures from ethnographic materials (Pink et al., 2022). This means using a design logic to develop ethnographic materials into design concepts through different iterations involving citizens, local stakeholders, public local strategists, and industry product developers. For instance, study four consisted of three steps. First, it tailored co-creation workshops with project stakeholders to transform ethnographic materials into intervention instruments. Second, it used these instruments to explore, imagine and design alternative AV futures. Third, it engaged again with the citizens to open further discussions and concerns about AVs in their daily lives. The first step was laying out ethnographic materials from the overarching AHA project and study three in an online collaboration platform, Miro. These materials were used to stage a collaborative engagement among user experience, service, interaction and business designers from an automotive company. The study utilised design fiction (Bleecker et al., 2022) to distance the participants from considering their everyday work boundaries or falling into discussing topics they are most familiar with and used to. The participants were asked to create AV concepts based on the ethnographic materials presented to them without concern with the feasibility of the concepts. Participants were asked to develop their future AV visions as speculative scenarios. To help them structure their scenarios, the study employed the common building blocks of what, when, why, where, and how used in scenario building (Carroll, 2003). The second step was held in situ with a broad set of project stakeholders. The participants were engaged in the speculative scenarios to trigger discussion and debate around AVs and to imagine alternative future AV scenarios. The scenarios were turned into videos and printouts to support such collabora-
tive engagement. This step elicited various societal concerns and anticipations of AVs that were then integrated into the speculative scenarios. The iterated scenarios were then used to engage with citizens, make meaning of the outcomes of the previous activities and take the discussion further. The result of this study was two-fold. First, it offered an insight into the different stakeholders’ and citizens’ perspectives on AVs. Second, it developed a tool that could be used to intervene with AV algorithm designers.

Study five employed speculative scenarios and semi-structured interviews to engage AV algorithm designers in collaborative, reflective practice (Amulya, 2011). The study involved algorithm designers from four companies working with AVs in workshops to interpret and make sense of the speculative scenarios. It employed previously developed speculative scenarios to trigger algorithm designers’ reflections on the impact of their work on society and their current work practices and organisational conditions that shape the AV algorithms. Semi-structured interviews allowed the structure of the workshops whilst allowing flexibility and opportunity for the participants to express what was important to them. This kind of engagement revealed organisational circumstances that limit opportunities for algorithm designers to be part of human-centred activities and the misalignment of stakeholders’ visions of AVs and the algorithm design realities.

These studies intentionally gathered different stakeholders for collaborative future-looking explorations (Pink et al., 2022), where knowledge was produced through collaborative meaning-making of the ethnographic materials (Pink & Morgan, 2013). The different engagements also informed the development of the DE methodology central to this thesis. Methodology that supports exploring the intersections between trust and AV development, trust and human-centred approaches, and AV development and human-centred approaches. It is also important to acknowledge that, in line with the notion of ethnographic place (Pink, 2009), the activities of the overarching AHA project contributed to the development of this methodology. As the research and other project activities comprised research sharing and collaboration with project stakeholders, the methodology gained from learning in all these situations cannot be attributed only to this thesis activities. Even more so, because of the long-term engagement through different research and project activities with project stakeholders and the problem area, AV development and future urban mobility provided insights not only into the social lives of citizens but also into the circumstances and perspectives of the stakeholders. This helped to tailor the workshops and the DE materials to support collaborative engagements but also to empathise with each other circumstances. In turn, this created collaborative learning from each other.
In conclusion, the five studies provided the primary data source for this thesis. At the same time, the overarching AHA project activities contributed to ethnographic long-term engagements with the city areas, as well as collaborative sense-making and knowledge creation necessary for developing and tailoring the DE methodology. The thesis data from the AHA activities were collected through audio and video recordings, digital materials created by the participants in the workshops, and researchers' notes on observations and reflections of the workshops. Table 1 presents an overview of the different data production points reported in this thesis.

Table 1. Overview of the studies that contributed with empirical data for this thesis.

<table>
<thead>
<tr>
<th>Study</th>
<th>Approach</th>
<th>Method</th>
<th>Participants</th>
</tr>
</thead>
</table>
| Study 1. Mapped the state-of-the-art in HCI trust in automation and AVs research. Identified a research gap. | Literature review.  
Case: HCI trust in automation and AVs research. | PRISMA Statement, A self-created tool. | - |
| Study 2. Explored how people’s real-life routines, practices and desires are accounted for in AV algorithm design. | Qualitative research approach.  
Case: AV algorithm design  
10 US professionals. |
| Study 3. Researched people’s trust in AVs in real-world context. | DE research approach.  
Case: Sharing AVs  
Analysis: a qualitative content analysis of the workshop documentation. | Co-Creation, Probing, Digital map activity. | 22 participants from the peri-urban area of Gothenburg, Sweden. |
| Study 4. Involved a heterogeneous set of AHA stakeholders to collaboratively explore people-centric materials and design speculative future AV scenarios. | DE research approach.  
Case: future mobility  
Analysis: continuously analysed the workshop documentation according to ethnographic analytic practice. | Co-creation, Design fiction, Scenario Building, Semi-structured interviews. | 12 professionals (in first four workshops),  
29 industry, academic, city, and local stakeholders (in one workshop),  
7 residents (in the final two workshops). |
| Study 5. Used speculative scenarios to engage AV algorithm designers in reflective practice to discuss their current work practices and possible effects AVs can have on society. | DE research approach.  
Case: AV algorithm design  
Analysis: continuously analysed the workshop documentation according to ethnographic analytic practice. | Co-creation, Reflective practice, Semi-structured interviews. | 13 professionals. |
Data Analysis

The data analysis methodology employed in this thesis adheres to the principles of ethnographic analytic practice outlined by O’Reilly (2012). This entails a concurrent and iterative data collection, sorting, summarisation, and organisation process. Each research endeavour and project undertaking serves as a foundation for subsequent activities within and across the various studies reported in the papers included in this thesis. This encompasses decisions related to the arrangement of research materials, the structuring of workshops, the selection of topics for exploration, and the engagement with individuals. A significant aspect influencing the data analysis pertains to the stimuli originating from internal and external sources, as highlighted by Merriam and Tisdell (2015). These stimuli, originating from within and beyond the scope of the research activities, play a crucial role in shaping the analytical process. As this thesis is part of larger research agendas in the TIC and AHA projects, the interactions and conversations with other researchers and stakeholders also inform, inspire and guide the interpretation and understanding of the collected data.

In the individual study contexts, despite the data analysis being done simultaneously with data gathering, the analysis is not finished once the last bit of data is collected. Instead, it becomes more intensive (Merriam & Tisdell, 2015) as the initial codes and themes get revised, and new ones are identified, having a sense of all the collected data. The analysis of the concrete ethnographic engagements is inspired by reflexive thematic analysis (Braun & Clarke, 2022; Terry & Hayfield, 2021). This has been tailored into the following steps:

1. Familiarisation with Data: The analysis begins with immersing in the collected data. This includes reading field notes, transcriptions, interview transcripts, and any other materials generated during fieldwork. It also implies gaining a comprehensive understanding of the context, participants, and their interactions.
2. Open Coding: Open coding involves generating initial descriptive labels (codes) for different elements or segments of the data. These codes capture vital concepts, actions, themes, and relationships.
3. Categorising: This step involves organising and connecting the initial codes into broader categories or themes. This step seeks to identify relationships and connections between different concepts and how they relate.
4. Creating central themes: This involves identifying the core themes that emerge from the data. These core themes are often the most significant and pervasive aspects of the phenomenon being studied.
5. **Constant Comparison:** The coding and categorising process engages in constant comparison. This involves comparing data within and across codes and categories to refine and validate the emerging themes. It helps ensure that the analysis remains grounded in the data.

6. **Contextualisation:** The findings are then situated within the broader socio-cultural, historical, and theoretical context and how the findings relate to existing literature and theories.

7. **Reporting through thick descriptions:** The final analysis is reported through rich and detailed descriptions that capture the nuances and complexities of the phenomenon under study.

8. **Ethical considerations:** The categories and themes are ethically validated by ongoing discussions of the findings with other researchers and collaborators in the project to reflect on how the analysis is situated from different stakeholders' perspectives (this step is described more in detail in the next section).

The analysis was supported by computer-assisted qualitative data analysis software ATLAS.ti. ATLAS.ti offers a broad set of features to manage and analyse qualitative data. This thesis utilised only a basic set of these features to organise the field notes, interview transcripts, and audio and video recordings to code the data and have an overview of the codes and themes that emerged through the analysis. ATLAS.ti also helped to manage and move in and out and between data from different studies.

### Research Quality and Ethics

It is a good practice to continuously evaluate the quality of qualitative research and if the conduct is ethically sound (Braun & Clarke, 2013) through various standards that qualitative research aspires to uphold. The research that underpins this thesis has been quality assessed through criteria proposed by Sarah J. Tracy (2010). She proposes a protocol of eight quality criteria: **worthy topic, rich rigour, sincerity, credibility, resonance, significant contribution, meaningful coherence and ethics.** The following section will elaborate on each of these criteria in relation to this thesis research.

Tracy (2010) defined qualitative research to have a **worthy topic** if it is relevant, timely, significant, interesting, or evocative. This thesis explored trust in the context of AVs. Both these areas are timely. Trust has been considered as one of the success factors for AVs (Wintersberger & Riener, 2016). This is especially as AVs are anticipated to bring societal benefits...
like increasing traffic safety, providing mobility for non-drivers, and decreasing the environmental impact of vehicles (Litman, 2021). As people's and public trust in AVs seems to be lingering on a low level (Maslej et al., 2023) while the industry is still pursuing the agenda, it is crucial to understand what is hindering people's trust in AVs and how it could be developed. In addition, AV development is an essential topic for the automotive industry. It wants to succeed in developing AVs with the anticipated societal benefits in mind (see Litman, 2021). However, the race to do that impacts society and people's lives. Therefore, exploring how to make AV development more responsible and trustworthy is also important.

Rich rigour can be understood as a researcher borrowing from various theories, collecting vast data, and allocating enough time, effort and care for the research (Tracy, 2010). This thesis research spans over four years, allowing engaging with trust and AV development over a long time from various perspectives and through different research activities. The empirical data is gathered in five studies, whereas other project activities also contribute to data analysis and method development. The research borrows from different well-known theories (see Baumer, 2017; Muir, 1994; Hoff & Bashir, 2015; Seaver, 2017; Stilgoe et al., 2013) and tools and techniques (see Amulya, 2011; Bleecker et al., 2022) used in other technology design and development contexts.

This thesis aims to be sincere and describe its results transparently and honestly (Tracy, 2010). It acknowledges the researcher's bias, e.g., in research methodology selection and when interpreting data. These nuances are discussed both in the individual papers and the cover paper. Sincerity is essential in this thesis as it is part of industrial PhD research. The possible bias imposed by the industry has been considered when analysing research data and has been reported in the cover paper but has not been elaborated on in the individual research papers.

Credibility is another aspect connected to communicating research. Tracy (2010) argued that qualitative credibility can be achieved through thick descriptions, triangulation and partiality. This research has always aimed to provide thick descriptions based on the principles of ethnographic practice (Crabtree et al., 2012) to help the audience imagine themselves in the research and participant situations (Pink, 2009). Therefore, the research process, situations, and circumstances have been described in detail in the individual research papers and the cover paper. Also, data gathering has followed qualitative research practices that ensure that produced results have credibility, e.g., reviewing research briefs and analysing data with research partners.
An essential aspect of research quality is its meaningful impact and resonance with an audience (Tracy, 2010). This thesis research lies between industry and academia. This duality imposes the research speaking to the scientific community and practitioners. A central concern in this research has been to tailor DE materials for different audiences (see Raats et al., 2023 (Paper 4) and Raats, 2024 (Paper 5)). As mentioned earlier, this research has produced research papers. However, the work has also involved activities like research dissemination in panel discussions, project meetings and the development of the AHA methodology catalogue (https://aha2.hh.se/wp-content/uploads/2022/11/AHA_methodology_catalogue_2022.pdf) with mixed audiences.

Research significantly contributes when it extends knowledge or improves practice through theoretical, practical, or methodological findings (Tracy, 2010). This thesis defines trustworthy algorithm design and proposes four elements that support AV development in shifting from technology-centred to human-centred practice and becoming more trustworthy. It adds to human-centred and algorithm design with methodological and theoretical insights. In addition, it contributes to HCI trust in AVs research by identifying a research gap (Raats et al., 2020) (Paper 1). The contributions of this thesis are the result of five studies that interconnect relevant “research design, data collection, and analysis with their theoretical framework and situational goals” (Tracy, 2010, p. 848) that are suitable for fulfilling the respected study goals and providing answers for research questions. Tracy (ibid.) calls this meaningful coherence.

One of the most important criteria of research quality is ethics. Tracy (2010) distinguished procedural, situational, relational and exiting ethics. This research borrows from Tracy’s (2010) eight quality criteria for qualitative research and good research practice principles provided by the Swedish Research Council (‘Good Research Practice’, 2017) to ensure and assess that good research standards are met. For instance, participant data like interview transcripts and video and audio recordings are managed appropriately. Also, the participants are treated and later represented respectfully. An essential ethical aspect is directly linked to the underlying DE methodology. Assuming DE as an interventional practice implies ethics of responsibility (Pink et al., 2022), undertaking DE research is ethical in its own right. This resonates well with the current thesis, as the research contribution calls for shifting towards more human-centred approaches that would nurture more trustworthy and responsible development to reduce the harmful impact of intelligent technologies. In addition to the thesis contributions written in this cover paper, these research findings have also been commu-
nicated in different academic and non-academic events like city expos, panel discussions, stakeholder workshops, research seminars and conferences. As with this cover paper, the main purpose has been to advocate for more responsible practices that focus on solving societal problems.
Summary of Papers

This chapter summarises the five papers that directly contribute to answering the overarching research question, “How can human-centred approaches support the development of trustworthy intelligent vehicle technology?”. Paper 1 maps HCI trust in automation and AVs and identifies the lack of insight into how people’s trust in AVs develops in real-world environments. Paper 2 aims to fill that gap and reports on socio-technical elements that mediate trust when people envision AVs as part of their everyday life environments and circumstances. Paper 3 investigates how AV algorithm design accounts for people-centric perspectives. Paper 4 engages citizens, city, academic, and industry stakeholders in DE people-centric materials to explore and discuss alternative future AV scenarios collaboratively. Paper 5

Figure 8. Overview of the papers and how they relate to each other.
uses DE people-centric materials to explore further how human-centred approaches are part of AV algorithm design and to study how algorithm designers can have more opportunities to partake in human-centred activities and responsible development. All the papers are described through the background, aim, method, and findings. See also Figure 8 for how the papers relate.

All the papers have been internationally peer-reviewed. One paper has been published in an international journal, two have been accepted in international conferences, one is in a third round of review at an international journal, and one has been submitted to an international information systems conference. For more details, attend to the whole research papers appended to the end of the thesis.

Paper 1


The first paper identifies a gap in HCI trust in automation and AV research. As AVs are a form of intelligent automation (Hengstler et al., 2016), the inquiry extends from AVs to include literature on trust in automation. This also allows learning from the insights on trust in automation. The study explores:

1. what are the key themes in HCI methodologies used to research trust in automation and AVs;
2. how do they account for trust in AVs as part of wider contexts; and
3. how can these methodologies be developed to include more than momentary and individual human-machine interactions?

To identify key themes in trust in automation and AVs research, this paper emphasises studies that focused on trust as the primary research question or investigated trust within a more extensive research agenda. Research reporting implications on trust as a by-product of another research focus is disregarded as they do not contribute to exploring trust in transportation research (see Figure 9).

The PRISMA Statement (Moher et al., 2009) defines a systematic analysis process which guides the literature review. The inquiry is made in five
academic research databases using common keywords identified in existing
HCI trust in automation and AVs literature. Using multiple databases max-
imises the data coverage and allows a more comprehensive understanding of
trust in automation and AV research. The search results in a total of 1709
records (565 from Engineering Village, 336 IEEE Xplore, 31 ScienceDirect,
522 Scopus, 255 Web of Science) that are narrowed down to 71 theoretical
and 258 empirical reports that are then analysed (see Figure 9).

Data analysis is supported by a self-made database interface (see Figure 10)
that allows, for example, tagging, categorisation and filtering. The interface
evolves throughout the analysis as the understanding of the literature and
trust phenomenon increases. For instance, a selection button is added to help
filter between relevant and irrelevant empirical and theoretical studies and
distinguish research in automation, automotive, and robotics.

The literature review results in findings on several levels. The findings
reveal many learned, dispositional and situational factors that influence trust
development (for example, see Hoff and Bashir’s full model of trust, 2015).
However, HCI trust in automation and AV research considers that trust develops in momentary and individual interactions between a person and technology. This has led to most trust studies being conducted in laboratory setups, test tracks and simulated scenarios and trust being assessed through already established questionnaires, e.g., the Trust in Automated Systems scale by Jian et al. (2000). For example, from the 258 empirical studies 221 assess trust using a questionnaire and only one explore trust through an ethnographic approach. This has an impact on what trust-influencing aspects are mainly being studied. The findings strongly focus on technology-related factors like design features and system performance.

![Image of literature analysis interface](image-url)

**Figure 10. A screenshot of the final interface used in literature analysis.**

**Paper 2**

This paper extends previous understandings of trust and explores trust in AVs in a social context they would be part of. It explores how trust develops when people experience technology in everyday real-life environments and social contexts (Hassenzahl, 2010).

Paper 1 contextualises AVs as part of a broader socio-technical system comprising various drivers, vehicles, cyclists, pedestrians, institutions, and rules and regulations (Henschke, 2020). More precisely, this paper considers AVs in the context of sharing, as sharing is an area where AVs are speculated to bring the most value, e.g., increasing the adoption of car-sharing services, decreasing the number of vehicles on the roads and reducing carbon emission (Hanna et al., 2016). The study investigates sharing AVs in concrete local communities, peoples’ existing relationships and everyday life circumstances. The study explores AVs and trust from a socio-technical system perspective.

This paper fills the gap in HCI trust in AVs research, identified in Paper 1, by investigating trust in shared AVs. Situating AVs as part of a more extensive socio-technical system allows for exploring how trust emerges when participants envision AVs as part of their daily life circumstances. The paper engages 22 participants, families coordinating their daily lives with each other in a peri-urban, semi-rural area of Gothenburg. Being part of a more extensive DE approach (Pink et al., 2022), the paper utilises co-creation workshops and probing to explore trust in participants’ existing and potential relationships and situations of the shared AVs. Paper reports on using

![Figure 11. Example of a digital map activity result.](image)
co-creation to involve various stakeholders and people in collaborative value creation (Ind & Coates, 2013). Probing engages workshop participants in collaboratively exploring and envisioning future autonomous mobility based on their present-day real-life circumstances and shared experiences (Çerçi et al., 2021). To comply with pandemic restrictions, seven online workshops are conducted utilising Zoom, Microsoft Teams or Google Meetup video conference system. The mural collaborative work platform facilitates joint activities around a digital map of the participants’ area (See Figure 11 for the digital map results from one of the workshops).

The paper reports on two-fold findings. First, situating AVs in a concrete local context helps to reveal aspects of trust not solely related to the underlying AV technology. More precisely, the importance of the social and physical environment in people’s anticipated experiences with AVs. Second, studying trust in shared AVs uncovers three important socio-technical aspects influencing trust in AVs. First, social trust in people’s anticipations of AVs is perhaps even more critical in trust in AVs than the underlying technical aspects (see Paper 1). Second, due to the absence of a driver, AV development should extend its focus to include solutions, e.g., geo and time blocking, operator services, and in-vehicle technologies that would facilitate safer AV sharing with strangers. Third, it prompts discussions about the opportunities AVs could bring. This is important, as it is common for people to accept technology based on their perceptions of the benefits the technology may offer them (Bronfman & Vázquez, 2011). However, the anticipated value is related to people’s daily routines, practices, and environments rather than the commonly envisioned long-term high-level benefits like increased traffic safety, mobility for non-drivers, and decreased environmental impact of vehicles (Litman, 2021).

Paper 3


This paper explores the AV algorithm design and how it accounts for people-centric perspectives. This is important as algorithms power intelligent technologies, e.g., AVs and are intended to learn and act differently in dif-
different situations. Furthermore, algorithm developers’ practices, attitudes, and assumptions influence the behaviour of algorithms and in real-world applications, algorithms often cause harmful situations (Allen et al., 2006), e.g., fatal accidents with AVs.

Scholars, therefore, call for a more human-centred algorithm design to limit these failures. One such contribution is made by Baumer (2017), who proposes the HCAD framework, which combines different initiatives under one umbrella and presents three strategies - theoretical, participatory, and speculative - to be applied in algorithm design to bring people’s perspectives into algorithm design and limit any harmful consequences of algorithmic systems. However, there is limited understanding of the organisational conditions that shape the AV algorithm design, which affects the algorithms implemented into the AVs. To fill this gap, this paper asks: “How are human-centred design strategies applied among developers of algorithms for autonomous vehicles?”

This paper is guided by Seaver’s (2017) theory of “algorithms as culture”, which focuses on how algorithms are enacted in practice and how algorithms become part of generating cultural traits. Ethnographic fieldwork (O’Reilly, 2012) is used to engage and learn from and about developers in their work environment at two offices of an automotive company. This approach identifies situations where human-centred approaches are practised, and human-centred perspectives are discussed to understand the algorithm design practice better.

The data is collected through in-depth semi-structured interviews with 29 AV developers - 19 in the main office and 10 in a satellite office. The developers mainly work with early-stage development projects, with some also occasionally in production. The developers working on early-phase projects are chosen because AI in the automotive industry is still being validated for its usefulness in proof-of-concept projects. The study consists of two groups and 15 individual interviews.

The interview recordings and fieldwork notes are transcribed and analysed through qualitative thematic analysis (Merriam & Grenier, 2019) using the coding software ATLAS.ti. The aim is to identify if, when, why, and how human-centred strategies are discussed and applied and to understand developers’ attitudes and assumptions guiding algorithm design for AVs. The study findings reveal three dominant levels in algorithm design practice: algorithm, social, and anticipatory (see Figure 12). The algorithm level consists of practices around data and data modelling. Data is crucial in algorithm development, and lacking data could risk a project being halted. Nevertheless, there is no systematic way to collect data. Data modelling depends on the designers’ conceptualisation of the world even though real-world
scenarios are categorised as “corner cases” and impossible to predict. The social level consists of team collaboration and demonstrating relevance in the organisation. Trust in colleagues and close cooperation are essential in algorithm design. However, a challenge is the distance and cultural differences between teams, requiring developers to prove their relevance constantly. On an anticipatory level, the developers imagine societal development needed to successfully release AVs into the real world and strategies so people would trust and start using the solutions in their daily lives. Overall, the algorithm designers acknowledge it is challenging for an individual actor, e.g., industry or government, to “solve” the AVs alone.

Figure 12. The three levels of algorithm development practice.

Paper 4

Research has shown that industry and policy-making envision future mobility as services and AI-driven technologies (Quilty et al., 2022). This means future mobility is part of a socio-technical system comprising, in addition to technological components, social, cultural, economic, and political aspects (Axsen & Sovacool, 2019). Failing to account for these aspects can lead to solutions that create unnecessary burdens to people and society instead of bringing value and being useful (Li & Voege, 2017). In turn, the often negative impact of technological innovation has been criticised by researchers who call for innovation to be more responsible (Frischmann & Selinger, 2018).

Therefore, this study gathers different societal actors to collaboratively explore questions of future autonomous mobility. Guided by Stilgoe et al.’s (2013) RI framework, it investigates how DE can be used to stage co-creation with citizens and city, academic, and industry stakeholders to foresee the possible unintended impact of future autonomous mobility on society. The paper asks: “How can ethnographic materials be used to tailor co-creation workshops to facilitate responsible innovation?”

This paper reports on a study carried out within the AHA project. AHA project grounded itself in a DE approach (Pink et al., 2022) involving various stakeholders in investigating future mobility in two neighbourhoods, areas suggested by project stakeholders from Gothenburg and Helsingborg city. These areas were engaged through ethnographic techniques (O’Reilly, 2012) to investigate local routines, practices and circumstances.

This paper describes how the ethnographic material is continuously tailored for upcoming co-creation workshops. It is done through seven co-creation workshops. The first four workshops are held online with ten designers from the automotive industry. Microsoft Teams and the online collaboration platform Miro help facilitate the workshops. The workshops consist of several steps of familiarisation with ethnographic material and a design task to develop a speculative narrative. These workshops offer new perspectives on people’s daily lives and anticipations of AVs. The workshops result in four speculative narratives of future autonomous mobility services. Speculation allows stakeholders to explore, discuss, and debate future mobility and where society might be headed (Bleecker et al., 2022) without their everyday work constraints. The next workshop is held in situ with 29 AHA stakeholders from the city, academia, industry, and local communities. Speculative narratives are used to prompt collaborative reflections and debate around issues of future mobility visions. The final two workshops engage a total of seven residents. Iterated speculative narratives are then used to open the created future mobility visions up for citizens’ critique and discussion on what value these can bring to their daily lives.
The workshops’ documentation is continuously analysed to tailor the co-creation process.


1. Tuning in the Everyday Life. Ethnographic materials guide the workshop set-up. For instance, the first four workshops are divided into familiarisation and design activities. Also, the materials are split into smaller parts, requiring the workshops to be held through five activities. During the first four activities, the participants were familiarised with insights about the areas, the practices and daily routines of people living there, descriptions of anticipated practical mobility solutions, and several concrete things residents had said during previously conducted workshops and fieldwork.

Starting with insights about peoples’ real-life environments, practices, and anticipations allows the participants to identify and define the problems. For many, this way is preferred as they were commonly involved in solving a list of problems without knowing where they originated or were prioritised. However, the challenge with introducing the ethnographic materials is making them graspable for the participants. This is where the workshops partially fail despite the materials being split into four parts.

2. Allowing to Relearn What is Obvious. In addition to the ethnographic insights, the workshop materials include descriptions of mobility solutions residents had envisioned to bring them value in their local areas and everyday life situations. As these offer new perspectives on what the participants are used to working with, the materials prompt collaborative reflections about their current work practices, attitudes, and assumptions about people’s daily lives. This indicates the importance of bringing such perspectives into the future mobility design process. Two challenges with this approach are allowing enough time for the participants to familiarise themselves with the material and communicating the insights in a graspable way. The workshops could do better in both aspects, provide more time and use visuals instead of only text to present the materials.

3. Creating Healthy Friction to Reach New Perspectives. Ethnographic materials can create friction by introducing new perspectives that challenge designers’ common attitudes and assumptions. These frictions are created by adding concerns and questions from residents and other stakeholders in different project activities. The speculation to add depends on the target audience; in the frame of these workshops, the focus is socio-technical trust issues. Overall, using speculative materials founded in peoples’ everyday lives supports exploring alternatives to existing mobility narratives and col-
laborative learning among the stakeholders. However, the challenge lies in balancing the speculations with realism to spark constructive debate.

4. Neutralising Opposition Between Stakeholders. A challenge in facilitating co-creation with, e.g., industry and cities is the seemingly conflicting expectations and agendas. Ethnographic people-centric materials offer a neutral ground for the stakeholders to meet, discuss and co-create. It also allows the stakeholders to get insights into each other’s practices and perspectives and overall trigger more engagement from the stakeholders. For instance, the stakeholders discuss city planning, business models, and equity in future mobility. They foresee the potential of co-creating preferred future mobility solutions instead of working on them individually.

Paper 5


This paper explores how reflective practice can advance HCAD and responsible AV development. It engages AV algorithm designers in DE people-centred materials to facilitate reflective practice and support developers’ opportunities to engage in value co-creation and responsible development. It is triggered by AV algorithm design mainly following a techno-solutionism logic and engineering visions, which do not necessarily align with people’s values and needs.

As the behaviour of algorithms depends on the actions, assumptions and attitudes of the algorithm designers (Allen et al., 2006), it has become even more critical to understand what aspects guide the algorithms. As a result, scholars have suggested more transparency in the development process (Mittelstadt, 2016) and accountability of the algorithm designers (Dia- kopoulos, 2014). There have also been calls for re-humanising algorithm development (Ruckenstein, 2022), and initiatives have been taken to bring insights from peoples’ social and cultural contexts into algorithm design (e.g., Forlano, 2019; Baumer, 2017). Baumer combined different human-centred initiatives under the HCAD umbrella to anticipate the unintended consequences of algorithms in a real-world context and align developers’ intentions of algorithms with people’s perceptions of their meaning and value.
The challenge, however, has been to operationalise this in technology development to involve different stakeholders to create value together with developers. This also requires more opportunities for algorithm designers to be part of such value co-creation and to engage them in ensuring algorithms do not cause damage in real-world applications. Value co-creation has been fundamental in service innovation (Lusch & Nambisan, 2015). The intention has been to switch from organisations developing value individually to involving customers in different parts of the business process to co-create value for society. Paper 4 explores how to facilitate RI principles (Stilgoe et al., 2013) through co-creation workshops. This paper investigates how reflective practice can expand algorithm designers’ opportunities to participate in value co-creation and responsible development. It asks, “How ethnoGraphically infused human-centred algorithm design can support reflective, collaborative practice?”

The paper reports on a qualitative research approach that explores how DE people-centred materials can trigger collaborative reflections on algorithm design practice. The use of materials is guided by Dourish’s (2006) calls for a broader use of ethnography than only for design implications.

The study involves algorithm designers from four companies and uses DE-infused ethnographic materials (see Paper 4 for how they were developed) to facilitate the discussions. The speculative nature of the materials aims to remove the algorithm designers from their everyday realities while reflecting on their current work practices and routines (Dunne & Raby, 2013). They will introduce people’s real-life habits, practices and expectations to complement the engineering visions currently dominating AV algorithm design.

The documentation created during the workshops is continuously analysed based on the ethnographic analytic practice (O’Reilly, 2012) and guidelines on staging qualitative research (Merriam & Tisdell, 2015). The documentation is coded by observing how the reflective practice appears through developers’ engagement with the DE materials. This leads to six themes of categories identified in the data through the analysis. How the developers engage with DE materials is guided by how HCAD can support reflective, collaborative practice (Amulya, 2011) in AV algorithm design.


1. Relationship with stakeholders. Algorithm designers value close collaboration with their stakeholders and find it essential for better quality and faster work results. DE materials also help investigate what other collabora-
tions must be set up to provide more holistic experiences and bring better societal value.

2. Current ways of working. Designers explain algorithm design to be driven from an engineering perspective. The focus is on safety, liability, and legal requirements. Also, organisationally, the designers are placed away from the end users and use context. Therefore, they have to rely on their peers to ensure that correct problems are presented for them to solve.

3. Algorithm development practice from a broader contextualised perspective. They, however, acknowledge the value of having a more comprehensive view of a problem. DE materials trigger interest in learning about different scenarios and ways AV could be used. However, in this paper’s context, they feel the scenarios provide a too limited view of what technology is capable of and what people truly need in their everyday lives.

4. Possible ways of knowing. Designers also reflect on how they currently learn, which tends to be through themselves or other engineers testing their solutions. If they are provided with user feedback, it mainly comes through complaint logs, rendering their task to solve issues rather than understanding a broader context of use.

5. Potential social and societal impact. The materials also prompt collaborative reflections on their algorithms’ possible effects on society and people. The designers speculate on how intelligent technologies must be introduced to be accepted and used by people. They explore potential issues such technologies could bring, e.g., data safety issues when different services and technologies interconnect and security issues with people sharing AVs with strangers—people-centred insights allow enriching engineering perspectives with real-life views.

6. Feasibility of the scenarios. The material challenges current assumptions and attitudes that guide AV algorithm design as people and stakeholders, e.g., user experience designers, anticipate AVs to be integrated into their lives more seamlessly than technologically possible. For instance, the participants point out that visions of pickups and drop-offs anywhere add significant complexity to AV development. Also, it is sometimes difficult for the designers to imagine how the presented material could directly contribute to their work tasks. Although, they find it helpful in initiating dialogue with stakeholders early in the problem definition stage.
Discussion

This thesis investigates the research question: “How can human-centred approaches support the development of trustworthy intelligent vehicle technology?” Through an empirical exploration of the intersection of research on trust, autonomous vehicle development and human-centred approaches, this thesis suggests an answer to the research question by coining a novel and timely framework for intelligent vehicle development: trustworthy algorithm design (TAD) (see Figure 13). This framework integrates ideas and

Figure 13. Elements of Trustworthy Algorithm Design.

strategies from different well-known trajectories of research in the field of responsible and human-centred technology development: Human-Centred
Algorithm Design (Baumer, 2017), trust in intelligent technologies (Muir, 1994; Hoff & Bashir, 2015), algorithms as culture (Seaver, 2017) and Responsible Innovation (Stilgoe et al., 2013). In addition, TAD contributes with empirically founded results and suggestions of how to stage the actual trustworthy algorithm design as an evolving process of cross-sectoral and interdisciplinary co-learning, as well as what elements and practices are needed to orchestrate this process of co-creating and making of trustworthiness. The following section begins by defining TAD, its principles, elements and practices in more detail. The chapter finally reflects on the limitations of this research and suggests potential future research.

Trustworthy Algorithm Design

Trustworthy Algorithm Design (TAD) is a framework that relies on the idea that trustworthiness is construed as a process to engage in instead of merely a measurable outcome or an innate product characteristic. This definition is based on three empirically grounded principles for TAD:

1. AV algorithms are more than merely a technology to be coded; instead, a design network comprised of stakeholders, technologies, contexts and local practices.
2. The AV algorithm design networks are continually evolving since the stakeholders, technologies, contexts and local practices are ever-changing, redefining AVs as they intersect.
3. The AV algorithm design networks develop through continuous co-learning in the intersections across sectors, disciplines, technologies, and local practices over time.

Table 2. Overview of the four TAD elements and their criteria.

<table>
<thead>
<tr>
<th>TAD Element</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Responsive</td>
<td>-Seeks to produce and integrate understandings from real-world contexts where algorithms are to partake</td>
</tr>
<tr>
<td></td>
<td>-Aims to be human-centred rather than technology-centred practice</td>
</tr>
<tr>
<td>Interventional</td>
<td>-Challenges conventional attitudes in algorithm design practice</td>
</tr>
<tr>
<td></td>
<td>-Explores potential impact algorithms can have on society</td>
</tr>
<tr>
<td>Intentional</td>
<td>-Leads in solving societal problem instead of fulfilling commercial briefs</td>
</tr>
<tr>
<td></td>
<td>-Educate other societal actors in possibilities and limitations of algorithms</td>
</tr>
<tr>
<td>Transdisciplinary</td>
<td>-Collaborates cross disciplines and sectors and with the public</td>
</tr>
<tr>
<td></td>
<td>-Stages engagements for continuous learning and evolvement with stakeholders developing and affected by intelligent technologies</td>
</tr>
</tbody>
</table>
These principles are built on four elements (see Table 2) that fuel the TAD process, based on empirical research underpinning this thesis. To develop TAD, the design practice must be responsive, interventional, intentional and transdisciplinary. The following section will describe these elements, argue for their importance, and suggest practical implications.

TAD Element 1: Responsive

To be responsive in the context of trustworthy AV algorithm design means that algorithm design practice is sensitive and adaptable to changing societal needs. It seeks to produce and integrate understandings of real-world contexts where the solutions are to partake into algorithm designs. It is a human-centred process that puts people first rather than a technology-centred practice.

As uncovered in Paper 3, there is a strong engineering focus in AV development. The main goal is to figure out how to solve the technological challenges of AVs. This implies an underlying technology-centred logic where the goal is to get people to use technology without considering whether the AVs fulfil their actual needs, an issue also pointed out by Baggozzi (2007) and Benbasat and Barki (2007). Paper 5 reveals that AV development primarily focuses on technological safety and liability to prevent people in vehicles from getting hurt. Little or no attention is given to social elements like the safety of sharing AVs with strangers. Although such elements are crucial in mediating people’s trust in AVs, as reported in Paper 2. Paper 5 also reports on the AV algorithm designers’ reflections on circumstances that hinder responsibility in their current work practices. AV algorithm designers are organisationally far from the end users, requiring them to assume aspects of the context they are developing for. Although the organisations developing AVs consider the people involved in the development to represent diverse perspectives, they mainly represent engineering views. The limited opportunities they get to be involved in discussions on what real societal problems AVs can solve require them to trust their peers with bringing them the “right” problems to solve. The responsibility for such decisions lies somewhere else in the organisations. However, as Brown et al. (2023) argued, the significant challenge in AV development is not technological but social. Therefore, the problem with distancing AV algorithm designers from the use context is that the algorithms will be as good as the algorithm designers' knowledge. As such, the behaviour of the algorithms is shaped by the algorithm designers’ assumptions, attitudes and beliefs (Allen et al., 2006). This is why AV algorithm designers need to get more understanding and opportunities to be part of engagements where find-
ings from real-world contexts are being discussed and produced—something they also desire, as by the Paper 5 results.

**Practical implications**

This thesis demonstrates how DE can balance technology-centred AV development with insights from people’s real-life situations and anticipations of AVs. This approach allows AV algorithm designers to engage in insights from people’s everyday lives. For instance, Paper 4 reports how co-creation workshops engage citizens, city, industry and academic stakeholders in collaborative explorations of future AV scenarios. The DE people-centric materials help to move the discussions beyond AV technology. An example of that is the workshop participants discussing and debating topics like safety mechanisms for people feeling comfortable in sharing AVs with strangers, how the cities need to be re-designing to accommodate AVs picking people up and dropping them off anywhere, and if everyone will be able to afford using AV services. Paper 5 uses scenarios of AV use to reflect on where such knowledge could be helpful and be brought into AV development. This revealed that such insights could help the AV algorithm designers to work faster and offer better-quality results. Also, as one participant expresses, seeing their work represented in potential use scenarios and contexts gives it more meaning.

**TAD Element 2: Interventional**

To be interventional means that trustworthy AV algorithm design challenges conventional attitudes and assumptions that underlie algorithm design practice. It employs human-centred approaches to investigate potential damage algorithms can cause to society by systematically imagining alternative future AV scenarios.

Scholars have called for technology development to anticipate a solution's potential impact on society (Stilgoe et al., 2013). This is reported to be lacking in AV development as an engineering belief guiding the development is that AVs will lead to long-term benefits like increased traffic safety and reduced environmental impact of vehicles (Litman, 2021). These anticipated goals, however, are too abstract to offer guidance or prepare the AVs for real-world context and vice versa. Also, they do not seem to assume any harmful impact of AVs. The real-life context is generally too complex to consider in algorithm design for AVs (as reported in Paper 3). However, a typical attitude among AV algorithm developers is that if automation is being released incrementally, people will accept and start using it. These problematic logics can lead to dominantly adopted assumptions and result in
fatal accidents as AVs fail to cope with real-world situations. For example, assumptions that the first and last mile are a challenge and can be solved with automation (Brodersen et al., 2023). The focus on safety from an AV technology perspective also implies a belief that people are primarily concerned with the underlying AV technology. This belief is contested in Paper 2, highlighting aspects like sharing AVs with strangers to mediate trust, as much as, if not more so, than the technological capabilities of AVs.

Nevertheless, algorithm designers have the expertise to speculate on the risks of developing intelligent technologies like AVs. For instance, Paper 5 highlights the algorithm designers’ concerns with visions where AVs are part of ecosystems sharing data with music streaming and social media services. They point out this kind of solution to introduce potential data privacy risks. This implies that AV algorithm designers need to be involved more in imagining and designing such future visions.

**Practical Implications**

This thesis demonstrates how AV algorithm designers can be engaged in foreseeing the potential societal impact of their solutions. It uses DE people-centric materials to engage AV algorithm designers and a broader set of stakeholders to discuss and debate AV futures. Papers 4 and 5 use people-centric materials infused with speculative aspects based on stakeholders’ (including citizens) concerns and anticipations of AVs. The initial DE materials give insights into people’s everyday routines, practices and desires. These are then infused with speculative aspects extracted from workshops where industry, city, and academic stakeholders collaboratively engaged with these materials. This creates situations where the stakeholders explore alternative versions of AV futures, not those they are usually exposed to and without concern about their everyday work boundaries (Bleecker et al., 2022). This triggers discussions and debate on other areas of development that are not necessarily related to the development of the underlying AV technology, for example, as highlighted in Paper 2, 4 and 5, how to ensure the safety of people sharing AVs with strangers, whether everyone will be able to afford to use AV services, and solving data privacy issues often presented in visions where different data needy services are connected to offer seamless user experiences. DE materials in these studies helped stage collaborative exploration of AVs' potential impact on society and surface potential risks in visions of future AVs.
TAD Element 3: Intentional

Trustworthy AV algorithm design takes the lead in solving societal problems and educating societal actors on the possibilities and limitations of algorithms. It calibrates ambitions, helps to define and works on solving “right” societal problems. This implies that trustworthy algorithm design is not a service that responds to commercial briefs to fulfil others’ technological agendas.

Scholars have criticised that design practice has been subjected to acting on commercial briefs rather than leading in problem-solving (Fry, 2009). This has left the designers in the dark about how their solutions make or break the world. The same can be observed in AV algorithm design. The algorithm designers are considered experts in their field. Paper 5 reports that they are approached with data and asked to see if it can be used for something or with a problem to solve. Often, they get no feedback if their answers or solutions solve anything. Organisationally, they are places away from the use context. Organisational circumstances impact their work also in other aspects. Paper 3 illustrates how the AV algorithm designers need to prove their relevance by staging successful concept demonstrations. The designers have developed strategies for keeping up with managerial decisions to align their work. When they come up with a problem to work with, it is challenging, if not impossible, to get it into final products. Managerial decisions cancel many projects without prior notice or discussion. This impacts the developed algorithms as they tend to be moulded by organisational politics rather than societal needs. However, as revealed in Papers 3 and 5, the AV algorithm designers are experts in their field. They can help to decide where to use and where to skip using algorithms and to foresee the potential risks with added intelligence. Therefore, the element of intentionality suggests that AV algorithm design needs to shift from fulfilling abstract visions to taking the lead in solving people’s actual needs and societal problems. This thesis does not advocate for more technology-driven development but rather to change how algorithm designers engage in societal problem-solving activities. They should be involved as equal partners whose expertise is crucial in every process step rather than used to serve answers and develop technology for someone else’s visions.

Practical Implications

The understanding that AV algorithm design needs to shift from servicing others to taking the lead and becoming more intentional in solving societal problems is based on previous research and empirical findings of the organisational circumstances that shape AV algorithm design and the attitudes the
algorithm designers hold. Paper 5 reports on a study that engages AV algorithm designers in DE materials to better understand the AV algorithm design practice. More specifically, it investigates where people-centric perspectives can be introduced into AV algorithm design and how algorithm designers can be more involved in human-centred activities in future AV mobility. The following discussion is based on the overall insights from this research.

Paper 5 discovered that AV algorithm designers have limited opportunities to participate in human-centred activities and access to contextual knowledge of where their solution will be applied. This suggests that an obvious solution is to provide algorithm designers with more opportunities to participate in such activities. This calls for considering AV algorithm designers as equal partners, not only experts in their field. This is an essential aspect as algorithm designer’s beliefs, attitudes and assumptions (Allen et al., 2006) determine how algorithms behave, and they can also calibrate people’s expectations and ambitions by identifying potential risks with applying algorithm. Paper 3 reports that AV algorithm designers need to continuously prove their relevance, which impacts what shapes the resulting algorithms. However, AV algorithm designers have the power to influence the algorithms that will determine the capabilities of AVs in real-world contexts. Therefore, they need to be acknowledged and involved in solving the correct problems as partners and not only as experts who can offer a perspective.

TAD Element 4: Transdisciplinary

Trustworthy AV algorithm design is a transdisciplinary practice of relevant cross-sector stakeholders and the public. Transdisciplinarity has been defined as a research practice that takes real-world problems as a starting point, acknowledges context-dependencies related to these problems, and integrates knowledge from different domains (Barth et al., 2023). In the TAD context, transdisciplinarity is staging engagements for collaborative learning and evolving with stakeholders developing and affected by AVs.

Scholars have pointed out that people do not trust solely the underlying AV technology but a network of actors affected by or partaking in AV development (Muir, 1994). This means AVs' success depends on different stakeholders and the public working together. Collaborative practices are common in AV algorithm design. Paper 3 reveals that algorithm developers work closely together with their peers to evaluate each other’s work to reduce the risk of coding errors getting through and work not getting done due to someone’s absence. Paper 5 reports algorithm developers combining
expert perspectives when defining and solving problems. They consider establishing good relationships with their peers to build trust and to increase the probability of their work being acknowledged and accepted. However, these collaborations mainly happen within the organisations and engineering disciplines. As reported in Paper 3, the developers recognise that for AVs to become successfully integrated into real-world context, external factors like city infrastructure also must evolve, but consider these issues out of their responsibility. Some exceptions exist, like algorithm designers collaborating with external stakeholders on vehicle safety regulations. New laws and regulations are also needed to safeguard passengers and pedestrians in case of accidents. Paper 4 reports on stakeholders' concerns about AVs' impact on urban planning and public transportation. This means AVs cannot be developed in isolation as they will impact different parts of society. It also discovered that both city and industry stakeholders mainly apply quantitative approaches in their work. Therefore, for AV development to succeed, the different sectors must adopt human-centred approaches that complement or develop current work practices. This will potentially increase the likelihood that AVs will meet societal needs and that the society is better prepared for adopting AVs. The participants voiced that, to begin with, it is good to involve citizens, city, academic, and industry stakeholders in imagining and exploring future AV scenarios (reported in Paper 4).

**Practical Implications**

On a practical note, this thesis demonstrates a way for staging such collaborations where the stakeholders can learn and experience human-centred approaches. Paper 4 used co-creation workshops to engage citizens, academic, city, and industry stakeholders to explore future AV scenarios collaboratively. It used DE materials to stage collaborative learning of people’s everyday circumstances and environments. These insights are presented through speculative scenarios and allow exploring topics and questions the stakeholders do not get to work on within their daily work. These scenarios triggered discussions that moved beyond AV technology and spanned across topics on urban planning, business cases, social inclusion and safety. An important aspect of using people-centric materials was creating a neutral space for, e.g. city and industry stakeholders to gather around as no one’s agenda was dominating. Collaboratively learning and making sense of the DE materials exposed the stakeholders’ perspectives on AVs, their role in society and their impact on their work. It also built empathy between the workshop participants on each other’s circumstances and triggered reflections on attitudes, e.g. towards citizens. For instance, a group of participants discussing an AV scenario voiced - “people are not numbers”.
Paper 5 zooms into a particular aspect of AV development, the AV algorithm design. It demonstrated how DE materials trigger collaborative, reflective practice (Amulya, 2011) among algorithm designers. It reports on co-creation workshops that engage the designers in people-centric materials to investigate what drives AV development and whether, where and how peoples’ everyday routines, practices and anticipation of AVs are accounted for. It reveals, for instance, that the algorithm designers play a crucial role in realising AVs while being mostly dismissed from discussions about the algorithmic risks and which societal problems to prioritise. Furthermore, the developers expressed that integrating human-centred activities and people-centric perspectives in AV algorithm design is valuable as it adds meaning to the work and increases the quality of the outcomes.

Contribution and Implications
It is common to describe how research strives to contribute new knowledge in three types - theoretical, methodological and empirical contributions. This indicates how knowledge comes from three different ways of knowing (e.g., Olson & Kellogg, 2014). Another typology developed in different fields within informatics also includes artefactual contributions (Wobbrock & Kientz, 2016; Ågerfalk & Karlsson, 2020). This thesis is primarily staged to give a contribution based on empirical research since it builds on ethnographic observations and interviews, as well as participatory modes of design. This empirical contribution has been theorised through two theoretical strands: a) understanding trust in intelligent vehicle technologies and b) human-centred intelligent technology collaborative design methodologies. The main contribution lies in the intersection of these theoretical strands, which have been brought together in this thesis to develop an empirically grounded understanding of algorithm design and how it can be staged to become trustworthy. In this way, this thesis contributes to research on trust in AVs by calling to focus on trustworthiness, as this will require a more comprehensive investigation into what mediates people’s trust. At the same time, it demonstrates how human-centred approaches like DE can stage explorations that situate AVs as not primarily a technology but as a constellation of places, people, and technologies that learn together in an evolving process in real-life contexts that mediate people’s trust.

This thesis mainly contributes by reimagining trust in AV development as a process of trustworthiness, AV algorithms as construed through a network of stakeholders, practices, and technologies, and therefore defining trustworthy algorithm design as a continuous collaborative learning and
evolvement process of different disciplines and sectors. It suggests that for AV algorithm design to be trustworthy, it needs to be responsive, interventional, intentional and transdisciplinary. What sets this contribution apart is that it calls to move away from considering trust as an attribute that can be developed into a technology and instead something that results from a trustworthiness process. This contribution implies to the industry that AV development needs to integrate human-centred approaches that help produce knowledge and understanding of the context the AVs will be part of. It needs to stage AV development to include collaborations with relevant external and internal stakeholders (and the public) to stage a process of co-learning and evolving together over time. It must acknowledge AVs as not merely a technology that can be manipulated to build people’s trust but as a collaborative effort that creates learning processes for stakeholders to participate in (including citizens) to ignite trustworthiness to evolve over time. In turn, this will also result in products people trust and that align with societal needs and, therefore, help the industry reach the anticipated societal benefits (see Litman, 2021). Through empirical studies, the thesis demonstrates how this can be achieved by employing DE, co-creation workshops and speculative approaches.

This thesis also contributes to research by identifying a gap in HCI trust in automation and AVs research literature (Paper 1). It shows a dominant application of quantitative research methods to research trust in AVs and a limited understanding of social aspects that mediate trust in AVs in real-world contexts. Therefore, this thesis also contributes by expanding the understanding of people’s trust in AVs with socio-technical elements that mediate trust when people envision AVs as part of their everyday lives. This implies that HCI trust in AVs research needs to employ qualitative techniques designed to capture the complexities of real-life context and extend the understanding of people’s trust. This thesis demonstrates a way of doing that through human-centred approaches like DE.

Another contribution this thesis makes is to the field of algorithm design. It uncovers various non-technology-related factors that influence the algorithms implemented into AVs. For instance, organisational structure and politics that distance algorithm designers from the use context and knowledge of people’s perspectives on AVs. The need for algorithm designers to prove their relevance, hence focusing on successful demonstrations of algorithmic systems rather than developing solutions to solve problems people face daily. This thesis also demonstrates how human-centred approaches like DE, co-creation and reflective practice can be applied to elicit these insights. These insights imply that for AV development to be more trustworthy, it needs to consider AV algorithm design not as a service
that enables the enhancement or development of new technology but as an equal partner in identifying and defining what societal problems need to be solved. In turn, this will help AV development also become more trustworthy.

This thesis also makes a methodological contribution and demonstrates how human-centred approaches and techniques like DE, design fiction, probing, co-creation, and reflective practice can be employed in different constellations to stage stakeholder engagements (including the public) and collaborative learning in AV development. It demonstrates how this moves the discussion away from the underlying AV technology, instead foregrounding societal aspects like access to AVs, the safety of sharing AVs with strangers and demands on city infrastructure. This shows the value of human-centred approaches in moving AV development from technology-centred to human-centred practice.

Limitations and Future Research

A limitation associated with qualitative approaches like DE, employed in this research, is the generalisability of the results (Merriam & Grenier, 2019). This is mainly attributed to the study scope and sample of participants. This notion also applies to the current thesis, which focuses on trust and the context of AV development and recruits mainly from Sweden, except for Study 2, where half of the participants resided in the US. This indicates two limitations. First, despite AVs being intelligent technologies, generalising the results to represent all technology development practices would be ill-fitted. Second, it is essential to acknowledge that the participants represent a narrow cultural and social context; hence, the findings also do not generalise to a broader population of developers and residents.

However, the goal of this thesis has not been to quantify trust and findings of AV development. Instead, this thesis offers knowledge and demonstrates an approach which can be transferred to other intelligent technology development contexts and when involving a (different) heterogeneous set of stakeholders. For example, Paper 1 reports the lack of social and cultural contextual knowledge on people’s trust in AVs and a bias towards quantitative approaches in HCI trust in AV research. This thesis contributes to filling that gap by using DE to elicit socio-technical elements mediating trust in AVs when people envision AVs as part of their everyday lives. Papers 3 and 5 revealed not only that AV (algorithm) development is mostly technology-centred, an insight that is not necessarily novel, but also circumstances why people-centric perspectives are dismissed in the development. Papers 4 and
5 disclose the value of using DE people-centric materials to balance engineering views and technology-centred perspectives with in-depth understandings of people’s lives and their anticipations of AVs.

Nevertheless, as Payne and Williams (2005) argue, qualitative research can also be generalised if the study meets specific requirements. This cover paper and the appended articles support this by providing detailed descriptions of the study contexts, approaches, methods, and discovered findings for further research to scrutinise and evaluate how much of it can be transferred.

That said, this thesis suggests future research to investigate how the elements of trustworthy algorithm design transfer to other algorithm design contexts, what other elements are crucial for algorithm design to be trustworthy and further explore how to support algorithm design from moving from service function to more intentional partaking in solving societal problems.
Concluding Remarks

This thesis explores the question, “*How can human-centred approaches support the development of trustworthy intelligent vehicle technology?*” It has shown why AV development needs to shift from focusing on trust as a technology attribute to trustworthiness as a continuous process. It demonstrates how that can be achieved through DE staging of collaborative learning and evolvement activities with stakeholders from different disciplines and sectors. Trustworthy AV development will increase people’s trust in AVs and help reach the anticipated societal benefits of AVs.

Intelligent technologies have great potential for society. However, the world seems to be going through an identity crisis, with rapid intelligent technology development without clear direction or purpose for how it can bring actual value to society, with new technology being tested at the expense of people’s lives, and with seemingly declining aspirations to have integrity and be trustworthy. This thesis supports intelligent technology development to find a direction and purpose and helps shift from technology-centred and engineering perspectives to more responsible, trustworthy and inclusive practice. I am glad to have had the opportunity to carry out this research. Thank you.
References


Amulya, J. (2011). *What is Reflective Practice?* The Center for Reflective Community Practice at MIT.


Ekman, F. (2023). *User trust here and now but not necessarily there and then—A Design Perspective on Appropriate Trust in Automated Vehicles (AVs)* [Doctor of Philosophy, Chalmers University of Technology]. https://research.chalmers.se/publication/534377/file/534377_Fulltext.pdf


Simon, J., & Rieder, G. (2021). Trusting the Corona-Warn-App? Contemplations on trust and trustworthiness at the intersection of technology, politics and public debate. *European Journal of Communication*, 026732312110283. [https://doi.org/10/gk8j7w](https://doi.org/10/gk8j7w)


Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy*, 42(9), 1568–1580. [https://doi.org/10/f5gv8h](https://doi.org/10/f5gv8h)


