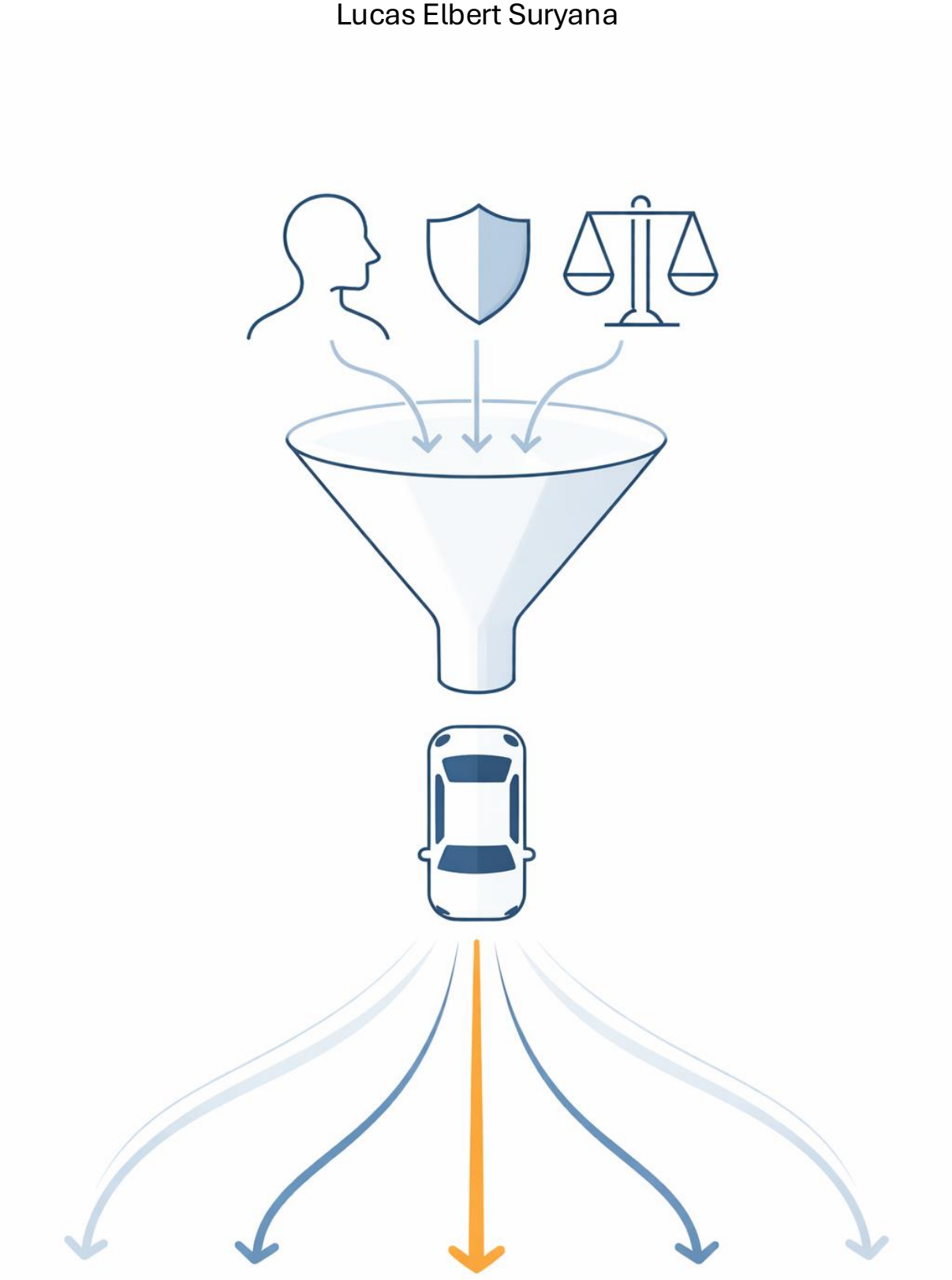
Operationalising Meaningful Human Control for Reason-Responsive Decision-Making in Automated Vehicles

Lucas Elbert Suryana



Propositions

accompanying the dissertation

Operationalising Meaningful Human Control for Reason-Responsive Decision-Making in Automated Vehicles

by

Lucas Elbert Suryana

Delft, 10 June 2026

1. Meaningful human control is not an end-state but a principle that should guide the design of automated-vehicle behaviour. [Chapter 2]
2. The design of automated-vehicle behaviour that responds to human reasons requires mechanisms to identify relevant human agents and to quantify the reasons underlying their actions. [Chapter 3]
3. Human-reason tracking must operate across all components of AV decision-making, from perception and prediction to evaluation and control. [Chapter 3 & 4]
4. The only way to keep a human operator feeling responsible during automated driving is to ensure they continuously perceive the situation as potentially hazardous. [Chapters 5 & 6]
5. One challenge in publishing research that operationalises meaningful human control in system decision-making is avoiding its framing as the sole primary contribution.
6. When writing a research article, you do not need to rush the submission; sometimes the insights that strengthen a paper emerge later as understanding deepens.
7. The statement 'This system is fully under meaningful human control' will never be heard.
8. You cannot fully understand a person by observing only their behaviour; yet many critical judgements rely on behaviour alone, overlooking the layers of underlying reasons that shape it.
9. A newborn, although fully human, is not yet meaningfully under human control.
10. Human beings naturally seek to express their intrinsic values beyond themselves, including through the artefacts they create.

These propositions are regarded as opposable and defendable, and have been approved as such by the promotors Prof. dr. ir. Bart van Arem and Dr. ir. Simeon Calvert and co-promotor Dr. Arkady Zgonnikov.

Operationalising Meaningful Human Control for Reason-Responsive Decision-Making in Automated Vehicles

Lucas Elbert Suryana

Operationalising Meaningful Human Control for Reason-Responsive Decision-Making in Automated Vehicles

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, Prof. dr. ir. H. Bijl
chair of the Board for Doctorates
to be defended publicly on
Wednesday 10 June 2026 at 12:30

by

Lucas Elbert Suryana

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	Chairperson
Prof. dr. ir. B. van Arem	Delft University of Technology, promotor
Dr. ir. S. C. Calvert	Delft University of Technology, promotor
Dr. A. Zgonnikov	Delft University of Technology, copromotor

Independent members:

Dr. N.E. Vellinga	University of Groningen, the Netherlands
Prof. dr. F. Santoni de Sio	Eindhoven University of Technology, the Netherlands
Prof. dr. J. Alonso-Mora	Delft University of Technology, the Netherlands
Prof. dr. M.H. Martens	Eindhoven University of Technology, the Netherlands

Reserve member:

Prof. dr. ir. R. Happee	Delft University of Technology, the Netherlands

This research was fully funded by Indonesia Endowment Fund for Education Agency.

Front & Back: designed by the author and generated with GPT-5.3.

TRAIL Thesis Series no. T2026/13, The Netherlands Research School TRAIL

TRAIL
P.O. BOX 5017
2600 GA Delft
The Netherlands
E-mail: info@rsTRAIL.nl

ISBN: 978-90-5584-388-6

Printed in the Netherlands

The sanctity of an oath with regard to promises, agreements, and contracts, has always been held in the greatest esteem, in every age and among every people.

Hugo Grotius
On the Law of War and Peace (1625)

Contents

Summary

Automated vehicles (AVs) are expected to improve road safety, efficiency, and accessibility, yet their behaviour can at times appear overly cautious, rigid, or counter-intuitive, undermining trust and public acceptance. Existing approaches to address this problem—ranging from ethical decision-making models to behaviour imitation and interaction-based design—often lack a principled account of *why* certain behaviours should occur in specific contexts. This dissertation argues that these limitations stem from the absence of a unified framework that links human reasons to automated-vehicle decision-making in a transparent and evaluable manner.

To address this challenge, the thesis adopts the philosophical framework of Meaningful Human Control (MHC), which requires that automated systems both track relevant human reasons and allow responsibility for outcomes to be meaningfully traced to human agents. While MHC has been widely discussed at a conceptual level, its technical operationalisation in automated driving remains underdeveloped. This dissertation advances MHC by translating its normative principles into an integrated framework that connects ethical reasoning, engineering implementation, and empirical evaluation.

The dissertation first investigates which human reasons are relevant for automated-vehicle manoeuvre planning in ethically ambiguous, everyday traffic situations. Empirical findings from interviews with AV experts show that such reasons are inherently multi-layered, context-dependent, and often simultaneous, spanning normative, strategic, tactical, and operational considerations. Rather than functioning as fixed values or isolated cost terms, human reasons are shown to form context-sensitive relationships between underlying motivations and expected vehicle behaviour. These insights provide an empirically grounded basis for structuring and prioritising human reasons in automated-vehicle decision-making.

Building on this foundation, the dissertation develops a technical approach for embedding human reasons within automated-vehicle control architectures. Human reasons are translated into formal, machine-readable representations by drawing on insights from human-factors research and are integrated through a supervisory evaluation layer that operates alongside existing motion planning and control frameworks. This approach enables transparent trajectory evaluation and adaptive behavioural adjustment without requiring the design of new controllers, thereby demonstrating a practical pathway for operationalising MHC in real-time decision-making systems.

Finally, the dissertation examines whether meaningful human control can be empirically assessed in practice. Qualitative studies with users of partially automated driving systems reveal how the tracking and tracing conditions of MHC manifest dynamically in drivers' experiences of safety, trust, responsibility, and intervention readiness. Complementary simulator experiments show that objective behavioural telemetry can capture aspects of tracking at the level of

concrete interaction events, while tracing cannot be inferred from behaviour alone. Together, these findings demonstrate that meaningful human control is not merely a normative or post-hoc concept, but an empirically observable property of ongoing human–automation interaction when evaluated through a multi-layer framework combining subjective perception and objective data.

Overall, this dissertation advances the technical operationalisation of meaningful human control by systematically linking human reasons, automated-vehicle decision-making, and empirical evaluation. The proposed framework provides researchers, designers, and policymakers with concrete tools to assess and support reason-aligned automated-vehicle behaviour, contributing to the development of automated driving systems whose behaviour is more transparent, context-sensitive, and reasonable in everyday traffic situations.

Samenvatting

Geautomatiseerde voertuigen (AV's) worden geacht de verkeersveiligheid, efficiëntie en toegankelijkheid te verbeteren. Toch kan hun gedrag in de praktijk soms overdreven voorzichtig, rigide of contra-intuïtief overkomen, wat het vertrouwen en de maatschappelijke acceptatie ondermijnt. Bestaande benaderingen om dit probleem aan te pakken—variërend van ethische besluitvormingsmodellen tot gedragsimitatie en interactiegebaseerd ontwerp—ontberen vaak een principiële onderbouwing van *waarom* bepaald gedrag in specifieke contexten wenselijk is. Dit proefschrift stelt dat deze beperkingen voortkomen uit het ontbreken van een samenhangend kader dat menselijke redenen op een transparante en evalueerbare manier verbindt met de besluitvorming van geautomatiseerde voertuigen.

Om deze uitdaging aan te pakken, maakt dit proefschrift gebruik van het filosofische raamwerk van Meaningful Human Control (MHC), dat vereist dat geautomatiseerde systemen enerzijds relevante menselijke redenen volgen en anderzijds toestaan dat verantwoordelijkheid voor uitkomsten betekenisvol kan worden herleid tot menselijke actoren. Hoewel MHC op conceptueel niveau uitgebreid is besproken, blijft de technische operationalisering ervan binnen geautomatiseerd rijden onderontwikkeld. Dit proefschrift draagt bij aan MHC door de normatieve principes ervan te vertalen naar een geïntegreerd raamwerk dat ethische overwegingen, technische implementatie en empirische evaluatie met elkaar verbindt.

Het proefschrift onderzoekt eerst welke menselijke redenen relevant zijn voor manoeuvreplanning van geautomatiseerde voertuigen in ethisch ambiguë, alledaagse verkeerssituaties. Empirische bevindingen uit interviews met AV-experts laten zien dat deze redenen inherent gelaagd, contextafhankelijk en vaak gelijktijdig zijn, en zich uitstrekken over normatieve, strategische, tactische en operationele overwegingen. In plaats van te functioneren als vaste waarden of geïsoleerde kostenfuncties, blijken menselijke redenen contextgevoelige relaties te vormen tussen onderliggende motieven en verwacht voertuiggedrag. Deze inzichten bieden een empirisch onderbouwde basis voor het structureren en prioriteren van menselijke redenen in de besluitvorming van geautomatiseerde voertuigen.

Voortbouwend op deze basis ontwikkelt het proefschrift een technische benadering om menselijke redenen te integreren in besturingsarchitecturen van geautomatiseerde voertuigen. Menselijke redenen worden vertaald naar formele, machinaal leesbare representaties, gebaseerd op inzichten uit het mens-factorenonderzoek, en geïntegreerd via een superviserende evaluatielaag die naast bestaande trajectplanning- en regelingskaders opereert. Deze benadering maakt transparante trajectevaluatie en adaptieve gedragsaanpassing mogelijk zonder dat nieuwe controllers hoeven te worden ontworpen, en toont daarmee een praktische route voor het operationaliseren van MHC in realtime besluitvormingssystemen.

Ten slotte onderzoekt het proefschrift of meaningful human control in de praktijk empirisch

kan worden geëvalueerd. Kwalitatieve studies met gebruikers van gedeeltelijk geautomatiseerde rijhulpsystemen laten zien hoe de tracking- en tracing-voorwaarden van MHC zich dynamisch manifesteren in ervaringen van veiligheid, vertrouwen, verantwoordelijkheid en interventiebereidheid. Aanvullende simulatorstudies tonen aan dat objectieve gedrags- en voertuigtelemetrie aspecten van tracking kan vastleggen op het niveau van concrete interactiegebeurtenissen, terwijl tracing niet uitsluitend uit gedrag kan worden afgeleid. Gezamenlijk laten deze bevindingen zien dat meaningful human control niet louter een normatief of post-hoc concept is, maar een empirisch waarneembare eigenschap van voortdurende mens–automatiseringsinteractie, mits geëvalueerd via een meerlagig raamwerk dat subjectieve perceptie en objectieve data combineert.

Samenvattend draagt dit proefschrift bij aan de technische operationalisering van meaningful human control door menselijke redenen, besluitvorming van geautomatiseerde voertuigen en empirische evaluatie systematisch met elkaar te verbinden. Het voorgestelde raamwerk biedt onderzoekers, ontwerpers en beleidsmakers concrete handvatten om reden-gealigneerd gedrag van geautomatiseerde voertuigen te beoordelen en te ondersteunen, en draagt daarmee bij aan de ontwikkeling van geautomatiseerde rijsystemen waarvan het gedrag transparanter, contextgevoeliger en redelijker is in alledaagse verkeerssituaties.

Chapter 1

Introduction

Automated vehicles (AVs) are increasingly expected to enhance road safety, mobility, and efficiency. However, in mixed-traffic environments where automated and human-driven vehicles share the road, AV behaviour shapes human trust and perceptions of safety. Technically safe but unnatural actions, such as overly cautious gaps or rigid adherence to rules, can deviate from what humans expect to be reasonable. Existing approaches remain high-level and conceptual, lacking an implementable method to enhance and enforce human reasons within decision algorithms. To address this challenge, this thesis aims to operationalise the concept of Meaningful Human Control (MHC) by formalising human reasons and embedding them into AV decision-making and evaluation.

This chapter begins by introducing the context and motivation for studying reasonable behaviour of AVs (Section 1.1) and defining the underlying research problem (Section 1.2). It then reviews existing approaches to human-aware AV design, including social, behavioural, and ethical frameworks (Section 1.3), and identifies their limitations. Section 1.4 presents meaningful human control as the conceptual foundation and explains current gaps in its operationalisation. Based on these discussions, Section 1.5 formulates the research gaps and overall objective, while Section 1.6 outlines the research questions addressed in this dissertation. The following sections describe the research approach and thesis structure (Section 1.7) and summarise the scientific and practical contributions of this work (Section 1.8).

1.1 Context and Motivation

Automated vehicles (AVs) are expected to transform modern mobility by improving safety, reducing congestion, and increasing accessibility (Agriesti et al., 2020; Matin & Dia, 2022). Yet these benefits depend not only on technical performance but also on how well AVs interact with human road users. In mixed-traffic environments, every automated action, such as accelerating, yielding, or merging, communicates intent to nearby people. When behaviour appears unnatural or counter-intuitive, it can undermine trust and acceptance, not only among surrounding road users but also within the wider public observing how AVs behave (Lee & See, 2004; Zhang et al., 2024).

Research has shown that when AVs adhere too rigidly to formalised rules of the road, their

behaviour can become overly cautious or insufficiently adaptable to the nuanced, judgement-based norms that characterise human driving (Bin-Nun et al., 2022). For instance, leaving large gaps or braking while the leading vehicle is still far ahead can disrupt traffic flow; maintaining the exact speed limit on a highway while others drive faster may increase safety risks; and following a cyclist for a long distance on a narrow two-way road can cause discomfort. Although such actions are technically safe, they can appear unreasonable, reducing comfort and potentially creating new safety concerns (Koopman & Widen, 2023). Furthermore, recent work by Cummings (2025) emphasises the importance of enabling AVs to reason under uncertainty, as reflected in the skill-, rule-, knowledge-, and expert-based (SRKE) taxonomy illustrated in Figure 1.1. The figure highlights the relationship between uncertainty and reasoning complexity: as uncertainty increases, more advanced reasoning capabilities becomes necessary.

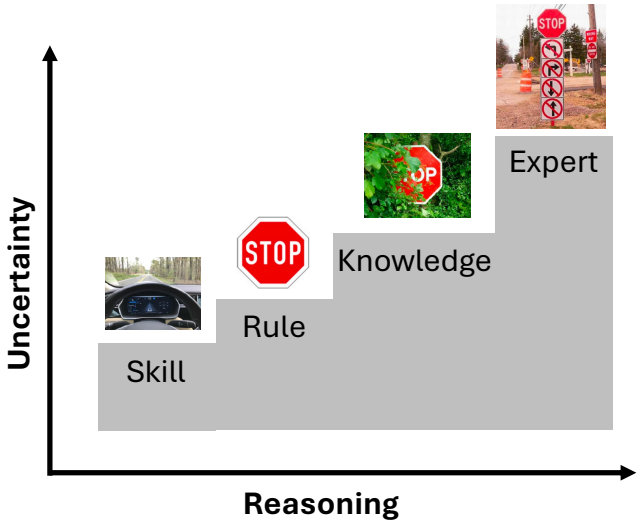


Figure 1.1: SRKE taxonomy (Cummings et al., 2024).

Thus, understanding and formalising what counts as **reasonable behaviour** is essential for designing human-centred AV decision-making systems. The next section defines this research problem and explains why existing approaches have not yet achieved it.

1.2 Problem Definition and Rationale

Automated vehicles with high levels of autonomy (e.g., SAE Level 4 systems (SAE International, 2021)) can already drive safely in many environments, including mixed-traffic urban settings. However, current control systems sometimes fail to respond as a human driver would in unexpected situations. For example, a recent accident involving a Cruise autonomous vehicle that dragged a pedestrian several metres (Quinn Emanuel Urquhart & Sullivan LLP, 2024) could be interpreted as illustrating the risks of rigid adherence to traffic rules. The vehicle's apparent assumption that no crossing was possible where none was legally designated, despite the common occurrence of jaywalking, may have contributed to the harmful outcome. This highlights that current control methods prioritise rule compliance while neglecting how humans actually behave around them. Such a mismatch between technical safety and reasonable behaviour threatens public trust and acceptance of AVs.

Existing studies have sought to bridge this gap through several directions, including models based on ethical decision-making, courtesy-driven behaviour, and social psychology (Thornton et al., 2016; Sun et al., 2018; Schwarting et al., 2019). Ethical decision-making frameworks encode philosophical principles, such as deontological duties or utilitarian harm minimisation, within model-predictive control (Thornton et al., 2018). However, they typically fix ethical priorities in advance and rarely justify how those trade-offs should adapt across different driving contexts. Courtesy-driven models, often trained using inverse reinforcement learning, generate behaviour that appears socially acceptable and comfortable (Huang et al., 2021), yet they infer such patterns from data without modelling the underlying reasons or justifications humans regard as courteous or fair. Social-psychology models draw on theories of interaction and behavioural prediction to simulate how road users influence one another (Zhang et al., 2023), but they typically reproduce mutual reactions without representing the normative reasoning or moral motives that guide those actions. Here, normative reasoning refers to the ability to justify behaviour in terms of shared human values and context-dependent social norms, rather than merely reproducing or optimising observed actions. Consequently, existing approaches replicate the appearance of reasonable behaviour without capturing the normative reasoning that makes such behaviour genuinely reasonable. Lacking this normative grounding, AV decisions can appear unpredictable or insensitive to context, deepening the trust and acceptance challenges described earlier.

Therefore, a unified framework is needed. This framework should be able to adapt across contexts and explicitly ground AV decision-making in the moral and practical reasons underlying human actions, ensuring that AV behaviour aligns with those reasons. The concept of meaningful human control has been proposed as a potential foundation (Mecacci & Santoni de Sio, 2020). MHC emphasises the need for automated systems to track relevant human reasons within their operational context and remain traceable to responsible agents. In this way, MHC offers a principled path toward embedding human reasoning and context-sensitive design into AV control. The following sections review related approaches and introduce MHC as the conceptual basis for this research.

1.3 Existing Approaches to Human-Aware AV Design

1.3.1 Social and Behavioural Modelling

Many studies have attempted to make automated-vehicle (AV) behaviour more human-like by modelling how drivers interact and cooperate in traffic. Typical approaches include imitation-learning frameworks (Sun et al., 2018; Jiang et al., 2023), game-theoretic interaction models (Schwarting et al., 2019; Liu et al., 2022), and social value orientation (SVO) formulations (Schwarting et al., 2019; Zhang et al., 2023) that represent altruistic or competitive motives. Imitation-based models reproduce trajectories observed in human data but can also replicate human errors, since they focus on behavioural replication rather than the underlying reasons for action. Game-theoretic frameworks treat driving as a strategic interaction, yet they rely on simplified mathematical utility functions that cannot easily represent human reasoning or moral judgement. Social-orientation models describe whether agents act prosocially or self-interestedly, but the use of such methods alone, such as SVO, merely reproduces human be-

haviour without capturing the reasons that make it reasonable. Although these approaches increase the realism of AV behaviour, they still reproduce only the surface patterns of human driving rather than the normative reasoning that justifies such behaviour as reasonable. Ethical and value-based frameworks have therefore been proposed to fill this gap.

1.3.2 Ethical and Value-Based Frameworks

Another line of research grounds automated-vehicle (AV) decisions in ethical and societal principles. These frameworks seek to translate moral values, such as duties, consequences, or virtues, into computational rules that can guide AV control (Thornton et al., 2016, 2018; Shea-Blymyer & Abbas, 2021; Geisslinger et al., 2023). Deontological models encode explicit moral duties, such as protecting vulnerable road users, but they struggle when those duties conflict with other values like rule compliance. Consequence-based or utilitarian approaches quantify harm and select actions that minimise expected loss, yet these depend on subjective weightings of whose risk matters more. Virtue-based ethics urge systems to act according to good character and relational care, but quantifying virtue and accounting for context-specific sensitivity remain major challenges. Although ethical frameworks provide a systematic foundation for considering human reasons, they often remain abstract, difficult to quantify, and detached from the situational reasoning humans apply in everyday driving. Addressing these limitations requires an approach that grounds ethical principles in the same cognitive and contextual reasoning humans use while driving, thereby linking moral abstraction to practical vehicle behaviour.

1.3.3 Limitations of Existing Approaches

Social-behavioural models make automated-vehicle (AV) behaviour appear more natural in interaction, while ethical frameworks explicitly introduce human values into decision making. Yet these research streams remain distant from achieving truly reasonable AV behaviour. Behavioural models reproduce what humans do on the surface but lack insight into why they act that way. Ethical frameworks, by contrast, encode why certain outcomes are valued, but remain abstract and detached from how those values should be applied in concrete driving situations. Each thus compensates for the other's weakness: behavioural models are realistic but normatively blind, whereas ethical models are principled but contextually thin. Their limitations are therefore interdependent, and addressing them in isolation cannot produce genuinely reasonable behaviour. What is needed is a unified framework that links human reasons with AV decision making while preserving behavioural realism. The following section introduces meaningful human control as one such concept.

1.4 Meaningful Human Control (MHC)

1.4.1 Concept and Origins

Meaningful human control is a philosophical framework developed to assess the extent to which humans retain control over the behaviour of automated systems, thereby keeping moral respon-

sibility with people rather than machines. The concept originated in debates on the ethics of autonomous weapons (Horowitz & Scharre, 2015) and was later expanded to address the "responsibility gap", known as situations in which no human can reasonably be held accountable for the actions of an automated system (Santoni de Sio & Van den Hoven, 2018). According to Santoni de Sio & Van den Hoven (2018), achieving MHC requires satisfying two core conditions: the **tracking condition** and the **tracing condition**. Tracking means that a system's behaviour should reflect the normative reasons of the humans who design, deploy, and interact with it. Tracing demands that the system's actions remain connected to identifiable human agents who understand how it operates, are capable of influencing it, and can be held responsible for its outcomes. Together, these conditions support the development of automated systems that can be guided by human reasons and remain under accountable human oversight, providing a foundation for analysing responsibility in automated-vehicle design.

To illustrate these conditions in the context of automated driving, consider an automated vehicle approaching a pedestrian crossing. Under the **tracking condition**, the vehicle's decision to slow down or yield should reflect relevant human reasons, such as the moral obligation to avoid harm and the intention to avoid collision. Under the **tracing condition**, there must remain an identifiable chain of responsibility for that decision, for example linking the vehicle's behaviour to the system designers who specified its decision logic or the driver who remains responsible for supervision where applicable. In contrast to common automated-vehicle control approaches that optimise predefined objectives such as safety or efficiency, meaningful human control explicitly requires that system behaviour remain responsive to human reasons, which may include but are not limited to safety and efficiency, and attributable to responsible human agents.

1.4.2 Application to AVs

Within the domain of automated driving, the tracking and tracing conditions of meaningful human control have been further specified to reflect the hierarchical structure of driving behaviour and the human–automation interaction characteristics of automated-vehicle systems. Mecacci & Santoni de Sio (2020) applied the concept of meaningful human control specifically to automated vehicles (AVs). They expanded the tracking condition by distinguishing four levels of human reasons—normative, strategic, tactical, and operational—adapted from Michon (1985) hierarchical model of driver behaviour. In their framework, normative reasons reflect motivations grounded in moral values (e.g., protecting human life); strategic reasons concern higher-level driving goals (e.g., route selection, risk tolerance); tactical reasons guide short-term manoeuvring decisions (e.g., gap acceptance, overtaking); and operational reasons involve immediate control actions (e.g., steering, braking). They also refined the tracing condition by emphasising that at least one human must remain within the vehicle's control loop, possessing sufficient understanding of system functions, the capacity to intervene when necessary, and awareness of moral and legal responsibility. Through these refinements, Mecacci & Santoni de Sio (2020) translated MHC from an abstract moral concept into a practical framework for designing and evaluating human reasoning and accountability in human–automation interaction within AVs.

1.4.3 Relation to Ethical Value Operationalisation in AV Behaviour

The framework developed by Mecacci & Santoni de Sio (2020) sits alongside a number of parallel and related studies that also seek to move beyond abstract normative principles toward operationalisable methodologies for aligning AV behaviour with ethical values (Bonnefon et al., 2020; Umbrello & Yampolskiy, 2022; Gros et al., 2025a). These efforts go beyond the frameworks reviewed in Section 1.3.2, which translate moral principles into mathematical structures such as costs, constraints, and formal logic, but do not provide a systematic methodology for determining which values should inform those structures or how societal and stakeholder preferences should shape them. For example, this can be achieved through multi-stakeholder ethical recommendations spanning road safety, risk and dilemmas, data and algorithm ethics, and responsibility (Bonnefon et al., 2020), value-sensitive design approaches (Umbrello & Yampolskiy, 2022), and empirically based approaches that derive moral attributes from societal preferences through user studies and translate them into Ethical Goal Functions (EGFs) through discrete choice modeling (Gros et al., 2025a). These works share the same motivation as MHC research: making AV behaviour transparent, value-sensitive, and accountable to human expectations. However, the operationalisation of ethical decision-making in these works focuses on addressing the question of what values AVs should adhere to, seeking to align them with the preferences of society. MHC, on the other hand, focuses on whose reasons a system tracks and whether its behaviour remains attributable to identifiable human agents. Methodologically, current framework implementations establish ethical objectives before deployment without addressing the integration of human reasoning at the behavioural level required by tracking conditions. This makes existing work on ethical decision-making complementary to MHC rather than competing: operationalising ethical decision-making informs what values should be embedded in AV systems, while MHC provides a framework for ensuring that system behaviour remains responsive and accountable to humans during operation, which is the concern pursued by this thesis.

1.4.4 Gaps in the Technical Operationalisation of MHC in AV

Despite these complementary developments, existing work on meaningful human control over AVs remains primarily conceptual, focusing on philosophical foundations (Mecacci & Santoni de Sio, 2020), component-based descriptions of tracking and tracing for AVs (Calvert & Mecacci, 2020), and conceptual frameworks for operationalisation (de Sio et al., 2023). What is less explored is how MHC, particularly the tracking condition, can be technically operationalised in practice. The tracking condition requires that system behaviour respond appropriately to human reasons, yet it remains unclear how automated vehicles (AVs) can be designed to reliably and explicitly incorporate such responsiveness. Current approaches often categorise human reasons or propose abstract frameworks but stop short of translating these insights into simulation-ready models or low-level decision rules that can govern AV behaviour in real driving scenarios. This gap highlights the need for research that bridges the philosophical foundations of MHC with the requirements of AV decision making systems and evaluation.

1.5 Research Gaps and Objective

Sections 1.3 and 1.4 show that existing research has advanced human-aware automated-vehicle (AV) design and developed the conceptual basis for meaningful human control. Yet there remains no systematic method for translating these philosophical principles into operational models and measurable behaviours for AV decision-making. This thesis therefore addresses three **key research gaps**, adapted from de Sio et al. (2023)'s suggestion for future work on MHC operationalisation:

1. **Ethics gap** - the absence of formal representation of human reasons suitable for AV decision making.
2. **Engineering gap** - the lack of implementation mechanisms that embed these reasons within vehicle control frameworks.
3. **Evaluation gap** - the limited empirical evaluation of whether MHC is realized in AV practice.

The objective of this dissertation is to operationalise meaningful human control for automated-vehicle decision-making by formalising human reasons, integrating them into AV decision-making algorithms, and developing methods to evaluate the reasonableness of AV behaviour.

1.6 Research Questions

Based on the research gaps identified in Section 1.5, this study formulates a set of research questions to guide the operationalisation of meaningful human control in automated-vehicle decision-making. Each question addresses a specific gap concerning the ethics, engineering, and evaluation aspects of MHC in AV behaviour. These three aspects form the structural components of the MHC operationalisation framework and are illustrated as pillars in Figure 1.2.

Ethics aspect: prioritising and formalising human reasons

Meaningful human control (MHC) requires that automated-vehicle behaviour track the reasons of human agents. Existing philosophical and conceptual work has proposed that such reasons can be distinguished across multiple levels, including moral, strategic, tactical, and operational (Michon, 1985; Mecacci & Santoni de Sio, 2020), and that different human agents may contribute proximal and distal reasons (Calvert & Mecacci, 2020). However, while these frameworks provide a normative structure for thinking about human–automation interaction, they do not systematically specify which types of human reasons are relevant in concrete automated-driving manoeuvres, nor how such reasons should be prioritised when multiple reasons apply simultaneously. As a result, current accounts provide insufficient guidance for determining **reason relevance and priority** in manoeuvre-level decision making.

Further, for automated vehicles to act in accordance with human reasons, those reasons must be represented in formal, quantifiable terms that can be implemented within planning and

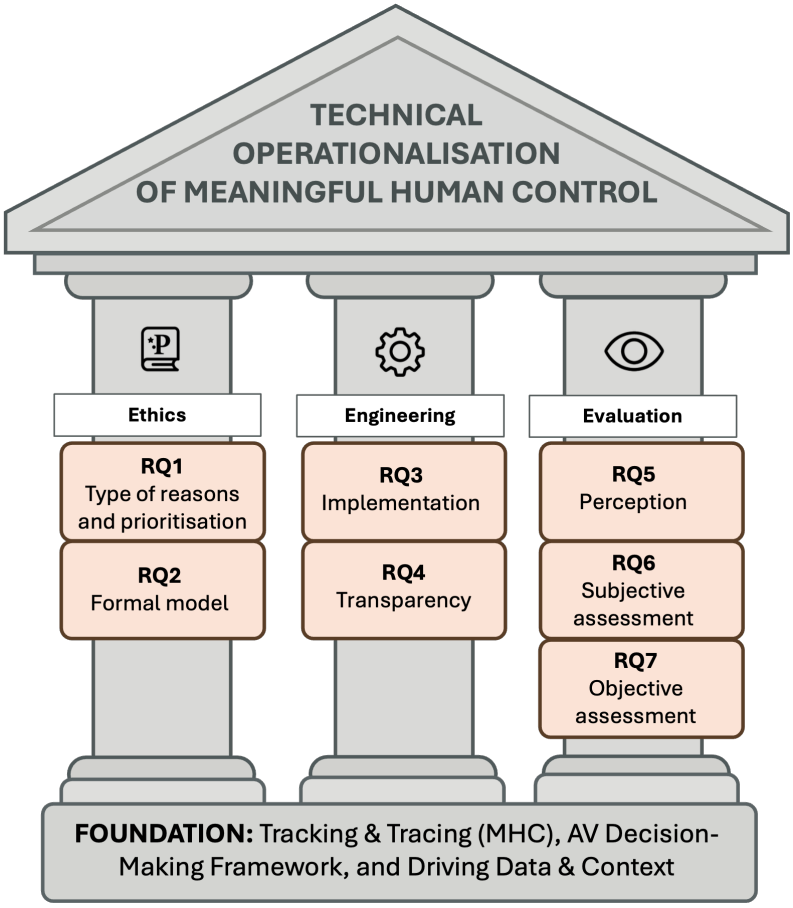


Figure 1.2: Pillar of MHC technical operationalisation.

control architectures. Although initial attempts have been made to express human reasons in formal models (e.g., Calvert & Mecacci, 2020), such representations remain hypothetical and rely on strong structural assumptions about how reasons map onto decision-making behaviour. As a result, it remains methodologically unclear how human reasons should be **formalised** in a way that is both computationally implementable and faithful to their normative role in automated vehicle decision-making.

Together, these gaps limit the operationalisation of MHC in automated-driving systems. They motivate the following research questions:

- **RQ1 (Type of reasons and prioritisation):** Which types of human reasons should automated vehicles consider and prioritise when planning manoeuvres?
- **RQ2 (Formal model):** How can these human reasons be formally represented in computational or mathematical terms suitable for automated-vehicle decision making?

Engineering aspect: embedding human reasons into control frameworks

Even if human reasons can be identified and quantified, they must still be embedded within automated-vehicle (AV) decision-making frameworks in order to influence actual vehicle behaviour. Some progress has been made in this direction, both in work that explicitly refers to

human reasons and in work that does not use this terminology but is nevertheless relevant to reasons-based control. With respect to the former, Calvert et al. (Calvert & Mecacci, 2020) propose embedding human reasons into control; however, this proposal is not integrated into the operational control frameworks that are standard in AV control research. With respect to the latter, there have been efforts to integrate ethical principles into model predictive control and trajectory planning (Thornton et al., 2016, 2018; Geisslinger et al., 2023).

Yet two issues remain. First, regarding explicit reasons-based control, **implementability** is limited: the proposal of Calvert & Mecacci (2020) remains at a conceptual level and is not integrated into widely used AV control architectures. Second, regarding value- and ethics-based control more broadly, **transparency** remains a concern: in a number of cases, the procedures for selecting, weighting, and trading off reasons or values are not made fully explicit, which constrains their auditability and accountability. As a result, there is currently no systematic and transparent method for evaluating whether AV trajectories align with specified human reasons, nor for adapting control behaviour when misalignment is detected. This motivates the following research questions:

- **RQ3 (Implementation):** How can vehicle control frameworks integrate human reasons to enable dynamic behavioural adaptation when misalignment occurs?
- **RQ4 (Transparency):** How can AV trajectories be systematically evaluated and selected based on relevant human reasons to satisfy the *tracking* condition of MHC?

Evaluation aspect: assessing tracking and tracing in practice

In addition to formalising and embedding human reasons into AV control frameworks, it is crucial to determine how to evaluate whether a system is operating under meaningful human control. Specifically, the system's ability to track human reasons and trace responsibility must be assessed. Without proper evaluation mechanisms, it is not possible to determine whether an AV is truly operating under MHC in practice. While prior studies have examined whether AVs align with tracking human reasons and tracing responsibility, most have relied on hypothetical scenarios or post-incident analyses (Calvert et al., 2020b, 2021). Although these approaches provide valuable insights, they do not capture the real-time, continuous evaluation needed to assess this alignment in dynamic, real-world driving scenarios.

Moreover, MHC is not only a technical issue but also a perceptual one, as it involves human interaction with AVs (Calvert et al., 2020a). The work describes that drivers' perceptions of safety and trust are influenced by how well the AV tracks their reasons (i.e., how well the AV's behaviour aligns with human motivations in a moral context). Additionally, drivers' understanding of their supervisory role is related to how traceability is maintained in the system's actions (i.e., how clearly they understand their responsibility). This simple scenario remains conceptual regarding the connection between human perception and MHC, but it is still unclear whether real drivers' **perceptions of safety and trust** truly correlate with MHC's tracking condition.

Furthermore, while perception plays a role in MHC, it remains uncertain whether current AVs can be evaluated under meaningful human control based solely on driver experience. **Subjective assessment** is important for evaluation because it reflects how drivers interpret and react to the AV's behaviour in real time. Such evaluation could provide insights into the alignment

between system design and human expectations. Another critical gap is whether behavioural telemetry data, including driver performance and interaction timing, can be used to objectively assess whether an AV operates under MHC in real time. **Objective assessment** is essential because it provides measurable evidence of whether the system adheres to MHC principles, beyond individual perception, and helps verify the consistency of the AV's behaviour across a range of situations. This complements the work of Verhagen et al. (2024), who underscore the importance of subjective experience in evaluating meaningful human control but also recognise the necessity of incorporating objective metrics to ensure that the system aligns with human expectations and operates effectively in real-world conditions. These gaps motivate the following research questions:

- **RQ5 (Perception):** How do drivers' perceptions of safety and trust relate to the extent to which automated vehicles track their reasons?
- **RQ6 (Subjective assessment):** How can drivers' subjective experiences be used to evaluate whether automated vehicles operate under meaningful human control?
- **RQ7 (Objective assessment):** How can behavioural telemetry data be used to evaluate whether automated vehicles operate under meaningful human control?

These questions collectively frame the methodological approach outlined in the next section.

1.7 Research Approach and Thesis Structure

Before outlining the research approach and the structure of the thesis, it is important to clarify the methodological scope. This work does not aim to design entirely new control algorithms but instead develops supervisory layers that operate on top of existing controllers. It also does not attempt to define ethical principles for every conceivable driving situation, but focuses on ethically challenging scenarios in which conflicting human reasons must be balanced. In the empirical analyses, the scope is narrowed to the human reason of safety, as evaluating all possible reasons simultaneously is not feasible. Finally, the thesis does not establish a ground-truth metric for MHC; rather, it investigates how metrics derived from telemetry data and subjective perceptions can indicate whether system behaviour aligns with meaningful human control.

To address the research questions in Section 1.6, this study adopts an interdisciplinary mixed-methods approach that integrates conceptual reasoning, computational modelling, and behavioural evaluation. These methodological strands correspond respectively to the **ethics**, **engineering**, and **evaluation** aspects of the research.

The **ethics aspect** establishes the philosophical and cognitive foundations of MHC and identifies the human reasons relevant to automated driving. This phase draws on expert interviews and qualitative analysis on their reasoning to develop a structured taxonomy of human reasons guiding AV behaviour. Building on these insights, the **engineering aspect** translates the identified human reasons into formal, machine-readable representations. Mathematical formulations grounded in human-factor principles are implemented as components of supervisory layers that monitor the alignment between AV behaviour and human reasons. Model Predictive Control (MPC) serves as the baseline control framework, while new mechanisms are developed

to detect and correct misalignment through reason-based evaluation and trajectory scoring. The **evaluation aspect** examines to what extent the tracking and tracing conditions of meaningful human control are reflected in both existing partially automated and simulated driving systems. This involves analysing data from the naturalistic driving experiences of users of commercially available partially automated vehicles, followed by the analysis of telemetry and qualitative data from controlled simulator experiments. Both qualitative and quantitative data are used to evaluate whether drivers' perceptions of safety, trust, and control correspond to measurable indicators in system telemetry.

The remainder of this dissertation is structured as shown in Figure 1.3. It consists of three main parts: the introduction, the main body, and the conclusion and perspectives. In line with the research questions, the main body comprises six chapters. Each chapter addresses one research question and is presented as a separate paper, except for Chapter 3, which addresses two research questions. Each chapter is based on a paper for which I am the first or joint-first author.

At the beginning of each chapter, the research question(s) addressed, the corresponding paper, and the publication status (published, under review, or submitted) are specified. This declaration is included because each paper was prepared for a different research field, requiring minor adjustments to individual introductions that may not fully align with the overall thesis introduction. These statements clarify how each chapter connects to the overarching research framework.

The six chapters are divided into two categories, tracking only and tracking and tracing, which are visually distinguished by blue and green colour areas in Figure 1.3. Black arrows illustrate conceptual relationships between chapters. **Chapter 2** investigates the reasons and principles for manoeuvre planning based on interviews with AV experts and a case study.

The contextual explanations developed in Chapter 2 are used in **Chapter 3** to formally represent human reasons as human-factor-grounded models (RQ2) and to implement a supervisory framework that evaluates the current AV state and triggers behavioural adjustments when misalignment is detected (RQ3). Building on the outputs of Chapters 2 and 3, **Chapter 4** focuses on scoring and selecting trajectories. Although Chapters 3 and 4 address different tracking mechanisms, both translate tracking-related concepts into conceptual implementations within automated driving systems (represented by the dashed grey arrow). The final component of the tracking-only category is **Chapter 5**, which investigates the relationship between drivers' perceptions of safety and trust in partially automated driving systems and system tracking.

Assuming that automated driving systems inherently involve both tracking and tracing, the second part of the main body extends the discussion to tracing. Building on the perception–tracking relationship identified in Chapter 5, **Chapter 6** further evaluates meaningful human control through subjective assessments. Finally, **Chapter 7** employs a driving simulator study with different driving control modes, using participants' simulated driving experiences combined with vehicle telemetry data, to provide an objective evaluation of meaningful human control.

The final chapter synthesises the key findings and presents the overall conclusions. Rather than repeating each chapter's results, it integrates them into a cohesive narrative and discusses their implications for science and practice, concluding with recommendations for future research.

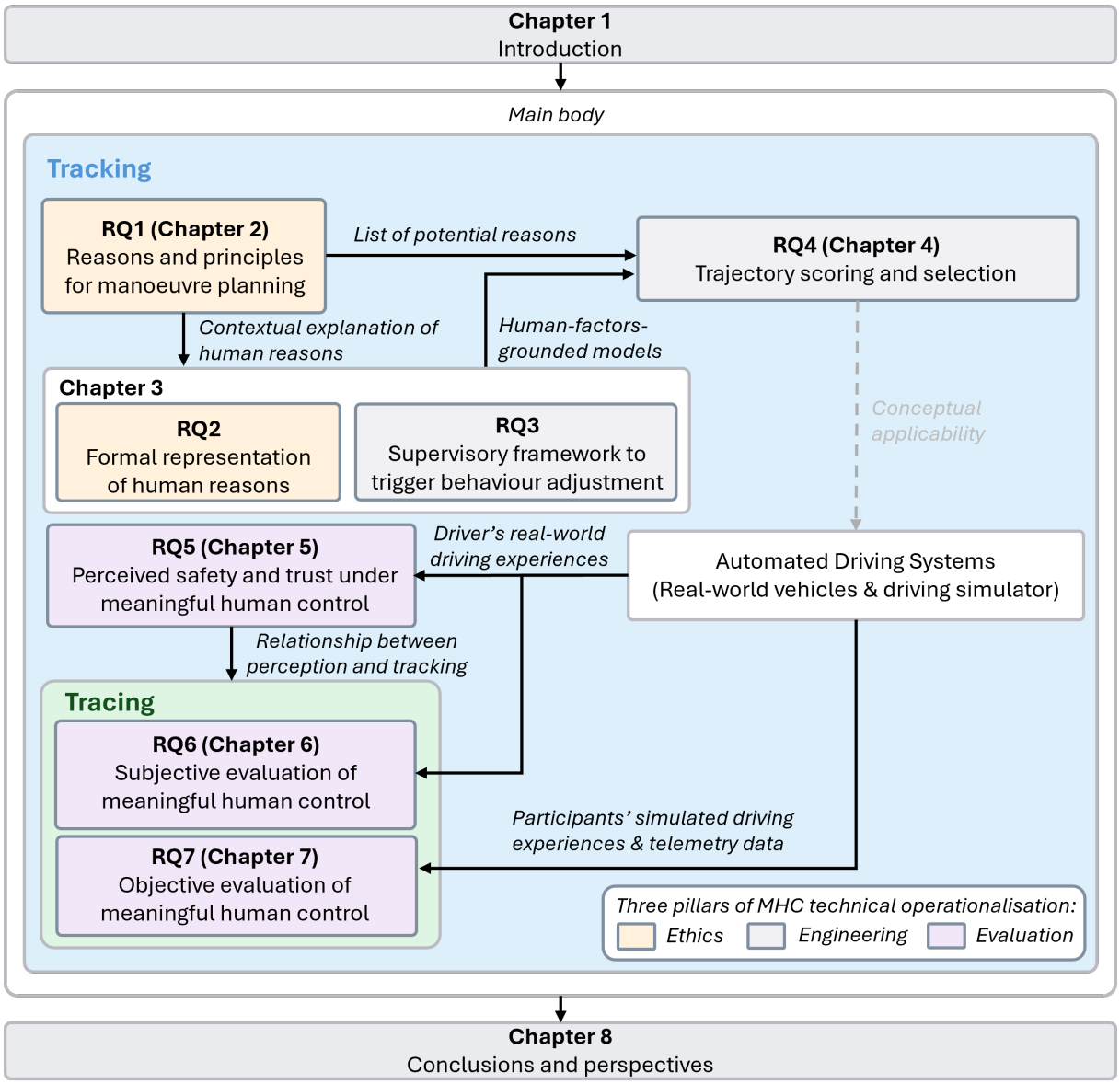


Figure 1.3: Outline of this thesis.

1.8 Scientific and Practical Contributions

Scientific Contributions

1. **Empirical grounding**: This thesis advances MHC research beyond conceptual discussions and simulation models toward empirical evaluation using real driving experience. Drawing on conversational data from Tesla drivers and controlled simulator studies, it demonstrates how drivers' perceptions of safety and trust systematically relate to the tracking and tracing conditions. This provides the first empirical grounding of MHC within partially automated driving systems.

2. **Operationalisation within control frameworks**: The thesis contributes to the implementation of MHC in practical vehicle control architectures. Whereas earlier work focused on taxonomies of reasons or high-level ethical models, this research embeds human

reasons directly into supervisory algorithms layered on top of Model Predictive Control (MPC). The proposed human-reason supervision framework enables vehicles to detect misalignment with relevant human reasons and trigger behavioural adjustments. In addition, a reason-based trajectory evaluation framework is introduced to assess and compare candidate trajectories based on human reasons, providing a systematic method for aligning AV behaviour with the tracking condition of MHC.

3. **Evaluation methodologies**: New methods are developed for assessing MHC that combine subjective perceptions and vehicle telemetry. The thesis pioneers the use of telemetry-level indicators as measurable proxies for the normative conditions of MHC, allowing systematic assessment of whether vehicles are, in practice, under meaningful human control.

4. **Advancing everyday AV ethics**: The research advances the growing shift in AV ethics from abstract trolley-problem scenarios toward ethically ambiguous yet everyday traffic situations. By analysing cases such as overtaking cyclists that balance rule compliance with safety, it demonstrates that the core challenge of operationalising MHC lies in managing ordinary but conflicting human reasons rather than rare thought experiments.

Practical Implications

1. **Conceptual support of regulatory assessment**: For regulators and authorities, the thesis provides a transparent, conceptually grounded evaluation framework for analysing the extent to which vehicles align with specified human reasons. The framework illustrates how policy-relevant values, such as prioritising the safety of vulnerable road users, could in principle be operationalised and assessed within automated-vehicle decision-making systems.

2. **Potential engineering integration**: For developers, the proposed supervisory frameworks are modular and lightweight in their conceptual design, suggesting that they could be integrated atop existing control algorithms without major architectural redesign. This indicates potential relevance for future industrial systems, subject to the requirements of safety validation and certification.

3. **Driving control design guidance**: The findings offer design principles for driving control systems, illustrating how different distributions of authority between driver and automation affect perceptions of agency and responsibility. These insights can help manufacturers maintain driver engagement while avoiding unrealistic supervisory demands.

4. **Transparency and accountability**: The proposed frameworks enhance the explainability of AV decision-making by quantifying alignment with human reasons. Making explicit why an automated system selects or revises a trajectory helps address regulatory and societal concerns regarding accountability and value alignment in automated driving.

Together, these contributions advance both the theoretical foundations and the practical realisation of meaningful human control in automated-vehicle research.

Chapter 2

Reasons and Principles for Automated Vehicle Manoeuvre Planning

---

This chapter investigates the human reasons and principles that guide automated vehicle (AV) manoeuvre planning in ethically ambiguous, everyday traffic situations. Building on the conceptual framework of Meaningful Human Control (MHC) introduced in Chapter 1, it addresses the question of which human reasons should be represented and prioritised when AVs make decisions in specific traffic scenarios. Using semi-structured interviews with AV experts from diverse disciplines and a targeted case study, this chapter identifies recurring categories of reasons and derives empirically grounded design principles for ethically aligned motion planning. The resulting framework provides the conceptual foundation for Chapter 3, which translates these qualitative insights into formal representations suitable for computational implementation.

This chapter is based on the following paper, published in *Transportation Research Interdisciplinary Perspectives (TRIP)*:

Suryana, L. E., Calvert, S., Zgonnikov, A., and van Arem, B. (2024). Reasons and Principles for Automated Vehicle Decisions in Ethically Ambiguous Everyday Scenarios: The Case of Cyclist Overtaking. TRIP, 35, 101787.

2.1 Abstract

Automated vehicles (AVs) consistently encounter ethically ambiguous situations in everyday driving, scenarios involving conflicting human interests and no clearly optimal course of action. While existing work often focuses on rare, high-stakes dilemmas (e.g., crash avoidance or trolley problems), routine decisions such as overtaking cyclists or navigating social interactions remain underexplored. This study addresses that gap by applying the tracking condition of Meaningful Human Control (MHC), which holds that AV behaviour should align with human reasons—the values, intentions, or expectations that justify actions. We conducted semi-structured interviews with 18 AV experts, who explained the reasons behind the considerations AV should make when planning a manoeuvre. Thirteen reason categories emerged, organised across normative, strategic, tactical, and operational levels. Using a case study on cyclist overtaking, we demonstrate how these reasons interact in practice and expose tensions in the decision-making process. Building on this analysis, we derive a reason-prioritisation principle grounded in the cyclist-overtaking scenario for AV behaviour in ethically ambiguous routine situations: prioritising vulnerable road users' safety above all, treating systemic safety and regulation as important but conditional, and permitting secondary values only when safety is not compromised. This hierarchy supports human-aligned behaviour by allowing pragmatic actions when strict legal compliance would undermine higher-priority values. Our findings offer conceptual principles intended to inform future research and design for AV decision-making in ethically challenging routine situations.

2.2 Introduction

Automated vehicle (AV) technology is advancing quickly, yet significant challenges remain, particularly when AVs must make decisions in ethically complex situations.(Nyholm & Smids, 2016; Wang et al., 2020; Saber et al., 2024). Such situations arise when AVs must balance multiple priorities such as safety, efficiency, and compliance with societal expectations, ranging from minimizing risks for all road users to resolving dilemmas involving conflicting values, interests, or trade-offs. Such conflicts often create ambiguity about the most appropriate course of action for AVs (Himmelreich, 2018; Bergmann, 2022). Addressing these dilemmas requires not only technical advancements, especially in planning and decision-making (Schwarting et al., 2018; Geisslinger et al., 2023), but also the development of clearer ethical principles.

2.2.1 Guidelines for AV in Ethically Straightforward Situations

Stakeholders involved in AV development and regulation have proposed various recommendations and guidelines for how AVs should behave in safety-critical situations. Among these, the concept of roadmanship has been introduced as a guiding principle to ensure that AVs drive safely, avoid creating hazards, and respond effectively to hazards caused by others (Fraade-Blanar et al., 2018). Although less formalised than regulatory guidelines, roadmanship emphasizes predictability and anticipatability in driving behaviour, similar to how a competent and careful human driver navigates traffic.

This concept aligns with the UNECE guidelines (UNECE, 2023), which present reference

models of "competent and careful" drivers. These models serve as benchmarks for evaluating AV behaviour in safety-critical scenarios such as cut-ins, cut-outs, and lead-vehicle deceleration, reflecting how a skilled human driver would minimise risk. If an AV outperforms these reference models, it is considered safer than a competent and careful human driver. Performance metrics include time headway, time-to-collision, and vehicle positioning, regardless of whether an accident is preventable (Olleja et al., 2025).

However, the safety-critical situations addressed by these models are ethically straightforward, since all parties benefit from the AV's behaviour. For example, the UNECE scenarios focus on mutually beneficial outcomes, such as avoiding collisions and reducing risk for all traffic participants. While these benchmarks offer important guidance for critical events, they do not address the full spectrum of ethically challenging situations that AVs may encounter in everyday driving.

2.2.2 Guidelines for AV in Ethically Ambiguous Situations

In everyday driving, AVs frequently encounter ambiguous situations involving conflicting interests, where the optimal course of action is unclear. Routine traffic scenarios—such as approaching a crosswalk with limited visibility or making a left turn in the presence of oncoming traffic—exemplify such dilemmas (Himmelreich, 2018). These situations often require balancing safety, legal compliance, and traffic efficiency. For instance, when approaching a crosswalk with limited visibility, an AV must decide whether to prioritise slowing down to ensure pedestrian safety, at the expense of reducing traffic flow, or maintain speed to optimise mobility, potentially increasing risk.

Unlike human drivers, who make such decisions intuitively on a case-by-case basis, drawing on experience and an understanding of human behaviour, AVs must encode these trade-offs systematically across all vehicles. This raises a fundamental ethical challenge: how should AVs navigate scenarios where competing priorities conflict at a systemic level?

Some researchers have examined these challenges through the lens of the trolley problem, which explores moral judgements in life-and-death trade-offs. One approach is rooted in consequentialist ethics, where actions are evaluated based on outcomes. For example, Meder et al. (2019) uncovered nuanced ways in which individuals' moral judgements reflect consequentialist reasoning. In contrast, deontological perspectives prioritise adherence to moral principles, such as the protection of passengers. Liu & Liu (2021) found that participants favoured AVs programmed to protect their occupants at all costs. Complementing these, the MIT Moral Machine project adopted a virtue-based approach, highlighting significant cultural variation in ethical preferences for AV decision-making (Awad et al., 2018).

While these studies reveal diverse ethical perspectives, they also underscore the difficulty of developing a universal framework for AV behaviour. Critics argue, however, that the trolley problem has limited relevance to real-world driving, as AVs are unlikely to encounter such stark life-and-death scenarios during routine operation (Nyholm & Smids, 2016; Keeling, 2020). Instead, AVs more commonly face ethically ambiguous situations where the stakes are lower but the complexity is higher—making these scenarios critical for AV development and deployment (Himmelreich, 2018).

Recent recommendations from a European Commission Expert Group propose guidelines

for addressing crash dilemmas through risk distribution and shared ethical principles (Bonnefon et al., 2020). Although these recommendations are valuable for ethically complex crash scenarios, they assume that crashes are unavoidable and do not fully address the challenges posed by routine, ethically ambiguous situations. This highlights the need for a more structured approach to navigating everyday ethical challenges, emphasising the importance of identifying principles that should guide AV behaviour in such contexts.

2.2.3 Guidelines based on the concept of Meaningful Human Control

The concept of Meaningful Human Control (MHC) emerged in response to concerns over wrongful actions by automated weapon systems, which could create a responsibility gap (Asaro, 2012; Horowitz & Scharre, 2015). Without meaningful human control, such systems might make life-and-death decisions that result in unintended casualties or violations of international law, such as targeting civilians rather than enemy combatants while the chain of responsibility is not clear. This concern also applies to AVs, where failure to resolve ethical dilemmas, such as balancing pedestrian safety with traffic flow, could result in safety-critical failures.

Santoni de Sio & Van den Hoven (2018) subsequently developed a foundational theory defining what MHC entails. According to this theory, automated systems should be responsive to the reasons provided by human agents for their decisions, ensuring alignment with human values and intentions—a principle referred to as the tracking condition of MHC. Researchers have since proposed ways to operationalise MHC for AVs. For instance, Mecacci & Santoni de Sio (2020) explores how the concept can be applied to the AV domain, while Calvert & Mecacci (2020) builds a conceptual model for implementing MHC in the control systems of connected and autonomous vehicles (CAVs) in mixed traffic environments.

We argue that the tracking condition of MHC provides a suitable foundation for developing guidelines to support AV behaviour in ethically ambiguous routine driving scenarios. Unlike rigid ethical frameworks based on predefined rules (e.g., utilitarian or consequentialist principles), MHC emphasises understanding and incorporating the reasoning behind human decisions. In such scenarios, human drivers often know how to act, relying on their ability to interpret context, weigh competing considerations, and make judgements accordingly. This natural adaptability underscores the importance of designing AVs that can dynamically respond to human reasoning in specific contexts. By focusing on the tracking of human reasons, MHC enables AVs to align their actions with human values and intent, an essential capability in real-world driving environments where ambiguity and unpredictability are common.

2.2.4 Research Gaps and Objectives

This study addresses three main gaps in current AV ethics research:

- **Theoretical gap**: While several AV ethics frameworks, such as the Integrative Ethical Decision-Making Framework (Rhim & Urban, 2021), have been proposed to address moral dilemmas, they primarily focus on high-stakes or exceptional crash scenarios. As a result, they often overlook the ethical complexity inherent in routine driving situations. The concept of MHC, particularly its tracking condition (Santoni de Sio & Van den

Hoven, 2018), offers a promising basis for addressing this gap by facilitating alignment between AV behaviour and human reasoning in everyday contexts. However, its application remains underexplored in ethically ambiguous routine scenarios. This study contributes by operationalising MHC to inform AV decision-making in such contexts.

- **Practical gap**: Although there is growing recognition that AVs must exhibit socially sensitive behaviour aligned with human values, expectations, and contextually appropriate responses (D'Amato et al., 2022; Lu et al., 2025), current frameworks lack a structured, empirically grounded set of human reasons for AVs to consider. This gap is especially evident in ethically complex, routine situations, such as overtaking a cyclist. This study addresses this gap by developing a principled framework that categorises expert-elicited reasons across normative, strategic, tactical, and operational levels of AV behaviour.
- **Methodological gap**: Existing research often relies on simulation models or moral vignettes, with limited empirical input from domain experts (Dubljević et al., 2023). While some studies incorporate expert perspectives in extreme scenarios (Milford et al., 2025), few systematically capture expert reasoning relevant to the daily ethically ambiguous conditions AVs encounter in everyday real-world operation. This study fills that gap through qualitative interviews with 18 AV experts, using a structured analysis informed by the MHC framework.

This study focuses specifically on perspectives from experts based in Western countries and should therefore be interpreted as reflecting Western views on AV decision-making. Rather than assuming universality, we aim to contribute an empirical framework for understanding the types of human reasons that AVs should consider in ethically ambiguous, routine driving scenarios. These region-specific insights form the basis for our methodological approach, which builds on the tracking condition of MHC to address the identified research gaps. Building on this foundation, the study has two primary objectives:

- **Objective 1**: To gather the reasons provided by AV experts regarding the factors they believe AVs should consider when planning a manoeuvre.
- **Objective 2**: To derive reason-based principles for AV decision-making by analysing the ethically ambiguous routine scenario of overtaking a cyclist using expert-derived reasons from Objective 1.

To address Objective 2, we adopt a two-step approach. In the first step, we classify the reasons elicited from AV experts into groups, aiming to identify underlying principles. In the second step, a case study of an ethically ambiguous routine scenario, such as overtaking a cyclist, is presented to the experts. They are asked to provide recommendations on the manoeuvre decisions AVs should make in such situations and to explain the reasoning behind their recommendations. These reasons are then mapped back to the classifications from the first step to uncover relationships and derive expert-informed guidelines for AV behaviour.

2.3 Reasons That Influence Considerations for AV Manoeuvre Planning

2.3.1 Methods

To identify the reasons automated vehicles (AVs) should track in ethically ambiguous routine driving situations, we conducted expert interviews. This section describes the selection of experts, recruitment procedures, interview protocol, and analytical approach used to extract and categorise these reasons.

Expert Participants

- *Selection Criteria*

  To ensure that the study reflected insights from individuals with substantive knowledge of AV systems, we employed a purposive sampling strategy, complemented by snowball sampling. Initial participants were selected through the professional networks of the authors and evaluated based on their publication records, institutional affiliations, and topic relevance, as reflected in publicly available sources such as Google Scholar profiles. This approach enabled the identification of experts with demonstrable contributions to AV-related research and development, in line with accepted practices in qualitative transportation studies. For example, Ma & Feng (2024) recruited AV professionals through LinkedIn based on their hands-on experience with automated systems, while Hilgarter & Granig (2020) employed purposive sampling in a real-world AV deployment by selecting participants immediately after they experienced an autonomous shuttle ride. Similarly, our approach aimed to ensure that participants had domain-specific expertise in AVs and were capable of contributing informed reasoning about AV decision-making.

  Subsequent participants were recruited via expert referrals following early interviews. This snowball sampling method enabled us to reach additional individuals working in specialised domains who may not have been immediately visible through conventional directories. Comparable combined strategies have been used in AV-focused qualitative studies to capture diverse, high-level perspectives from academia, industry, and government (Milford et al., 2025).

- *Recruitment*

  Participants were recruited via personalised email invitations. Each email included a brief overview of the study, highlighting its focus on understanding trade-offs in motion planning for overtaking scenarios involving automated vehicles (AVs). We clarified that although the study focused on motion planning, participation was not limited to specialists in that area; instead, we sought a broad range of perspectives from individuals involved in AV ethics, design, policy, and engineering.

  The email also outlined the interview format and logistics: semi-structured, approximately 45–60 minutes in duration, conducted via Zoom, and optionally recorded with participant consent. The voluntary nature of participation and the right to withdraw at any time were clearly communicated. Recruitment and study procedures were approved

by the Human Research Ethics Committee (HREC) of Delft University of Technology (ID: 132530).

- Participant Profile

  The final sample consisted of 18 expert participants from seven countries, representing a range of perspectives on AV development. Of the 35 experts initially contacted—14 from the United States, 13 from the Netherlands, four from the United Kingdom, and one each from Italy, Belgium, Israel, and Japan—18 agreed to participate, resulting in a response rate of 51%. All participants were informed of the study objectives and provided consent prior to their involvement.

  The participant profile reflects diverse institutional affiliations and technical backgrounds. As shown in Table 2.1, participants were drawn from academia (n = 12) and industry (n = 6), with disciplinary expertise in motion planning, human factors, ethics, behavioural science, and legal policy. Based on self-reported experience, participants had on average more than five years of direct involvement with automated vehicle development. This diversity of expertise and roles contributed to a rich and multidimensional set of perspectives on AV decision-making in motion planning contexts.

Table 2.1: Overview of Expert Participants by Sector, Country, and Expertise

ID	Country	Expertise	Role
Academia			
1	Netherlands	Human-AI interaction, ethics	Researcher
2	US	Technical validation, travel behaviour	Researcher
3	US	AV safety validation	Researcher
4	Netherlands	Motion planning algorithms	Researcher
5	Netherlands	Road users and infrastructure perspectives	Researcher
6	Netherlands	Ethics of AI	Researcher
7	UK	Modelling human behaviour	Researcher
8	Netherlands	Social science of behaviour	Researcher
9	UK	AV user experience	Researcher
10	Israel	Public perception and AV ethics	Researcher
11	UK	Human factors in transport	Researcher
12	Netherlands	Legal aspects of AV	Researcher
Industry			
13	UK	AV safety and assurance	Consultant
14	Netherlands	Traffic psychology	Psychologist
15	US	Software quality assurance	Engineer
16	US	Driving strategy, business development	Consultant
17	US	AV safety	Consultant
18	US	Human factors	Researcher

The sample size of 18 experts aligns with prior qualitative studies that employ in-depth expert interviews in the domains of automated vehicles, where 9 to 19 participants are often sufficient to achieve conceptual saturation (Dreger et al., 2020; Tabone et al., 2021;

Beringhoff et al., 2022; Lee et al., 2020; Swain et al., 2023; Habibullah et al., 2024). Although our sample size was determined by expert availability rather than a predefined saturation threshold, we conducted a retrospective assessment to evaluate whether thematic saturation was likely achieved. We tracked the emergence of new reason categories across interviews and observed that all categories described in Section 2.3.2 were identified by the 14th interview. The final four interviews introduced no new categories, suggesting that the major themes had stabilized. This provides additional confidence in the adequacy of the sample size within the scope and aims of this study.

Questionnaire Design

This study used a semi-structured interview protocol, operationalised through a structured questionnaire administered synchronously during interviews. The instrument was informed by the tracking condition of the Meaningful Human Control (MHC) framework. This condition holds that automated systems should respond to relevant human reasons (Santoni de Sio & Van den Hoven, 2018). In this article, we define "reasons" as normative reasons or factual considerations that justify particular actions, rather than motivational reasons, following the distinction outlined by Veluwenkamp (2022).

According to the MHC framework, reasons relevant to automated vehicle (AV) decision-making can be grouped into four categories: moral, strategic, tactical, and operational. Moral reasons pertain to ethical principles or social norms (e.g., fairness, harm avoidance). Strategic reasons relate to long-term planning goals (e.g., minimising travel time). Tactical reasons involve interactions with other road users (e.g., overtaking or yielding), while operational reasons concern real-time control actions (e.g., braking, steering).These categories provided the conceptual basis for the questionnaire, which consisted of five main parts:

- **Part 1 (Questions 2–4)**: Exploration factors that influence AV manoeuvre planning.
- **Part 2 (Questions 5-9)**: Evaluation of an ethically challenging real-world AV scenario involving a cyclist, asking participants to identify and assess reasons relevant to decision-making.
- **Part 3 (Questions 10-13)**: Ranking of predefined reasons, enabling participants to indicate the most appropriate decision in the given scenario.
- **Part 4 (Questions 14-19)**: Evaluation of alternative AV decisions, using time-based assessments to examine how stakeholder reasons were addressed in a revised scenario.
- **Part 5 (Questions 20-21)**: Evaluation of how AVs might interpret stakeholder intentions and manage potential conflicts.

The protocol was informally piloted with five PhD researchers working on topics related to AVs to ensure question clarity and relevance, after which wording adjustments were made prior to data collection. Building on this refined protocol, this section focuses on participants' responses to Questions 2–4, which relate to Objective 1 and are presented in Table 2.2. For completeness, the full set of questions is provided in Appendix A.1, while details of how both open- and closed-ended questions were formatted and administered are described in Section 2.3.1.

Table 2.2: Interview Questions Relevant to Objective 1

No.	Interview Question
Question 2	What should automated vehicles (AVs) consider when planning a maneuver? Please give one example in as much detail as possible.
Question 3	Which moral aspects do you believe AVs should consider when planning a maneuver?
Question 4	How might these aspects affect the maneuver plan?

Using these questions as the starting point, we elicited expert views on the kinds of reasons AVs should respond to. Participants answered open-ended questions designed to explore what factors should be considered in AV manoeuvre planning. Their responses to Questions 2–4 formed the basis for a theory-driven qualitative coding process aimed at identifying the types of reasons referenced and mapping them to the four categories outlined in the MHC framework. The analysis procedure is detailed in Section 2.3.1.

Procedure

The interviewer conducted all interviews online using Microsoft Teams, with audio, video, and automated transcriptions recorded for analysis. Each session followed a predefined protocol consisting of both open-ended and closed-ended questions presented through Qualtrics (https://www.qualtrics.com). This synchronous format , the interviewer opened the questionnaire on their own screen and shared it with participants via screen share. A direct, non-recorded link to the same questionnaire was also provided, enabling participants to reread questions and revisit previous items independently. This link also allowed them to rewatch embedded videos, which was especially helpful in cases of video lag caused by internet issues. The interviewer read each question aloud and asked participants to respond verbally. For closed-ended questions, the interviewer recorded participants' responses directly into the questionnaire, with the input visible to participants via screen share for confirmation. To prevent duplicate data, the interviewer explained that any responses submitted via the shared link would not be recorded or considered in the analysis. The interviewer also managed the structure and flow of each session. Participants were informed of this format in advance, and no concerns or discomfort were reported during or after the interviews.

This synchronous format enabled the researcher to provide immediate clarification when needed and ensured that participants responded to questions in the intended sequence. It also helped maintain consistency across interviews, as all participants saw and heard the same content in the same order at a similar pace. The researcher did not comment on or react to participants' responses and refrained from offering prompts or interpretations, intervening only when participants explicitly requested clarification. This neutral and minimal involvement allowed for observation of subtle cues, such as hesitation or clarification requests, that could enrich qualitative analysis. This method aligns with best practices for structured qualitative interviewing and has been applied in prior research (Longhurst & Johnston, 2023; Beringhoff et al., 2022; Nordhoff et al., 2023).

Data Analysis

- Coding Framework

  We used directed content analysis to analyse expert responses, using the Meaningful Human Control (MHC) framework (Mecacci & Santoni de Sio, 2020) as the initial coding structure. This framework distinguishes reasons according to their position on a temporal scale—that is, how close or distant they are from influencing an action—and organises them into four layers. These layers were used as deductive codes to classify the reasons experts provided for expected AV manoeuvre planning. The layers are detailed as follows:

  - **Normative reasons:** Motivations grounded in moral values, legal rules, or social expectations that guide what ought to be done. These are abstract, long-term in scope, and typically shaped by institutions or broader societal expectations.
  - **Strategic reasons:** Motivations or intentions related to high-level goals and long-term plans, such as deciding where to go or what outcome to achieve. These are moderately abstract, span longer durations, and are usually attributed to the driver as planner.
  - **Tactical reasons:** Motivations or intentions that guide short-term manoeuvring decisions in response to changing circumstances. These are more concrete and informed by the immediate driving context.
  - **Operational reasons:** Immediate motivations or intentions that correspond directly to moment-by-moment physical actions. These are highly specific and implemented by the AV system or human driver in response to moment-to-moment environmental cues.

  Based on the four layers of reasons, we created a coding matrix. Interview responses to Questions 2–4 were segmented into individual statements, which were then coded according to the type of category expressed. Once all statements were categorised, we conducted an inductive thematic analysis within each category to identify more specific sub-themes.

  This process enabled us to identify which layers of reasons were most frequently cited by experts and to characterise the diversity of reasons underlying what the AV should consider when planning a manoeuvre. For example, reasons in the moral layer often referred to legal compliance or fairness, whereas strategic reasons focused on acceptance and efficiency concerns. Tactical reasons addressed situational decision-making, and operational reasons emphasised vehicle control. When statements reflected more than one type of reason, cross-coding was used to preserve interpretive nuance.

  These four layers of reasons were used exclusively to code expert reasons. The subsequent analysis of what experts believed the AV should do, presented later in Section 2.2, was carried out separately as an exploratory interpretive step, using behavioural levels (normative, strategic, tactical, and operational) (Calvert & Mecacci, 2020) that were not part of the coding framework.

  Our approach, grounded in the theoretical framework of MHC, is consistent with other AV studies employing theory-driven content analysis. For instance, Aasvik et al. (2025) used a similar method to analyse public trust in autonomous shuttles, while Suryana et al.

(2025b) applied the MHC framework to explore interview data in relation to the tracking and tracing conditions.

- *Qualitative Content Analysis Procedure*

  To apply the Meaningful Human Control (MHC) framework in a structured and transparent manner, we conducted a qualitative content analysis that combined theory-driven and data-driven steps. We began by segmenting interview transcripts into individual response units. Each unit was analysed to identify four key components: (1) the AV behaviour being recommended (consideration), (2) the justification for that behaviour (reason), (3) the human agent associated with the reason, and (4) the corresponding layers of reasons: normative, strategic, tactical, or operational.

  We explicitly distinguished between considerations and reasons. "Considerations" refers to the specific behaviour that the AV is expected to perform (e.g., "the AV should slow down near pedestrians"), whereas "reasons" are the human-orientated justifications for those behaviours (e.g., "to ensure the safety of vulnerable road users"). Each reason was then evaluated according to the layers of reasons. This involved examining its temporal scale and the associated human agents to whom the reason was attributed. When a reason encompassed multiple types of justification, such as combining moral fairness with strategic efficiency, it was assigned to more than one MHC category.

  Our approach recognised that reasons could be either explicitly stated in the data or logically inferred from context. Explicit reasons were identified when participants directly articulated the justification for their statements. In other cases, implicit reasons were inferred based on the surrounding narrative. This approach draws on principles of latent content analysis, in which underlying meanings are interpreted beyond the literal language used. Latent content, as defined by Graneheim & Lundman (2004), refers to the deeper meaning embedded in a text, especially important when participants allude to motivations or norms without stating them directly. Building on this, scholars such as Vaismoradi et al. (2013) and Krippendorff (2018) have emphasised how interpreting latent content can uncover implicit yet meaningful patterns within qualitative interview data.

  Following the identification and classification of reasons, responses that addressed similar topics were grouped into broader thematic categories. This step enabled us to organise the data into a set of distinct reason types, each of which was then linked to the appropriate MHC category or categories.

- *Inter-coder Reliability*

  To ensure the reliability and transparency of the coding process, we adopted a multi-stage, consensus-based approach. First, the one of the author compiled an initial list of reasons or expectations for how AVs should act, based on expert responses. This involved interpreting each expert's response and identifying distinct reasons or expectations expressed.

  Drawing on the classification framework proposed by Mecacci & Santoni de Sio (2020), we categorized each reason or expectation into four categories: moral, strategic, tactical, and operational. Following this categorization, two authors of this paper independently coded each item to one of the four categories to ensure consistency and analytical rigor. The initial coding was done independently using the same list, allowing for a direct comparison of interpretations.

We calculated inter-coder agreement using Cohen's kappa, based on binary coding of whether each of the four categories was applied. Agreement varied by category and coder pair. For example, the **moral** category showed substantial agreement ($\kappa = 0.77$ for Coder 2 vs Coder 1), while the **tactical** category showed moderate to substantial agreement ($\kappa = 0.62$ for Coder 1 vs Coder 3). In contrast, agreement was lower for the **strategic** ($\kappa = 0.11$–0.22) and **operational** ($\kappa = 0.00$–0.18) categories, indicating greater interpretive variability in these dimensions. This aligns with recent work showing that annotator disagreement can itself be a signal of underlying subjectivity, especially when arguments are tied to human values (Homayounirad et al., 2025). Most discrepancies arose from ambiguous phrasing in participant responses or overlapping themes across categories (e.g., a reason could plausibly be interpreted as both moral and strategic). Operational justifications were particularly prone to divergent interpretation, likely due to their context-specific nature. These differences were discussed during a follow-up meeting until full consensus was reached, and no disagreements remained unresolved.

After reaching agreement at the category level, we collaboratively developed sub-categories within each of the four main categories, refining the framework through discussion. During this process, one co-author noted that some sub-categories could conceptually belong to more than one main category, depending on context and interpretation. These overlaps were acknowledged and addressed through further discussion, with final coding decisions made by consensus. This approach enhanced the clarity and consistency of the framework while minimizing individual bias. It also reflects established best practices for investigator triangulation and consensus coding in directed content analysis (Hsieh & Shannon, 2005; Hill et al., 2005; Campbell et al., 2013).

2.3.2 Results

In response to Questions 2–4, most experts described hypothetical traffic situations and outlined the considerations and actions that automated vehicles (AVs) should take before executing a manoeuvre in these contexts. They also provided explanations (reasons), articulating why the considerations they mentioned were important and why they believed AVs should exhibit particular behaviours. Based on thematic analysis, we identified thirteen distinct categories of reasons that span across four layers of reasons used by experts to justify AV behaviour. These categories demonstrate that experts' expectations about what the AV should consider or do rely on diverse human reasons. Some categories may seem similar when viewed only by name, as labels necessarily compress complex reasoning into short descriptors. Examining the temporal scale and the layer to which each reason is classified helps reveal their conceptual differences. To clarify differences in layer and temporal scale, we developed a conceptual representation that positions each reason category along two dimensions.

Figure 2.1 presents this conceptual space, showing how the thirteen reason categories distribute across their levels of reasons and their temporal scale, with colours indicating the relevant human agents referenced in each reason. The visual organisation highlights that distinctions among categories are more apparent when considering all dimensions together (layers of reasons, temporal scale, and the relevant human agents referenced) rather than relying solely on category labels.

Following this visual clarification, Table 2.3 summarises each reason category and shows

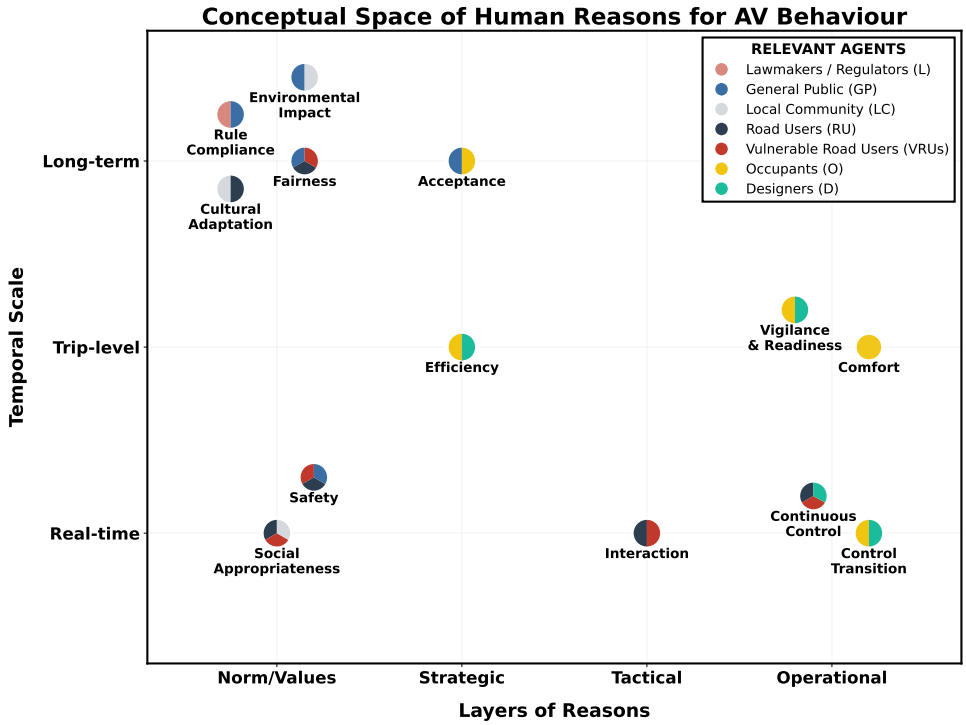


Figure 2.1: Conceptual positioning of the thirteen reason categories along two dimensions: layers of reasons (norm/values, strategic, tactical, and operational) and temporal scale (real-time, trip-level, and long-term). Relevant human agents are shown using colour, indicating the individuals or groups referenced in the justification for why the AV should behave in a particular way. This representation clarifies the conceptual focus of each category.

how we organised the associated reasons across four behavioural levels (normative, strategic, tactical, and operational), providing a clearer structure for understanding the types of actions experts described. We now describe each of the thirteen reason categories in turn.

Rule Compliance

Rule compliance refers to how the AV follows formal traffic laws and signals. Experts described this category as centred on legal duties that define how road users are expected to act by regulators and the general public.

One expert explained that the AV should follow traffic signals and road rules because these reflect shared legal norms that shape public expectations (ID 1). Another expert said the AV should obey traffic signals and signage to respect legal obligations and maintain rule-following behaviour in traffic (ID 3).

An expert also discussed how the AV should respond when dangerous rule violations occur. They suggested that the AV could warn nearby vehicles and record serious breaches, such as running a red light, so that responsibility can be assigned when needed (ID 16). This was described as supporting safety by providing clear information for those who enforce the rules.

Experts further commented on the connection between rule-following and responsible be-

haviour. One expert said the AV should follow road rules and act like a "good driver," meaning that rule compliance is part of demonstrating responsible conduct (ID 18). Another expert argued that the AV should follow traffic rules even when moral guidance is unclear, since rules remain a stable reference in those situations (ID 11).

A final contribution highlighted that strict compliance can reduce situations where the AV faces unclear or conflicting choices. One expert explained that following traffic law helps avoid ambiguity because the rules offer a widely accepted basis for deciding how to act (ID 12).

In summary, rule compliance was described by experts as relevant across multiple behavioural levels. At the normative level, rule compliance was described as the AV behaving in accordance with legal expectations established by lawmakers and society. At the tactical level, rule compliance appeared in discussions of visible behaviours such as obeying signals and signage, which other road users rely on to understand how the vehicle will act. At the operational level, experts emphasised that the AV must consistently execute rule-following actions, ensuring that its behaviour aligns with legal requirements in real time. This shows that rule compliance depends on how the AV connects widely accepted legal expectations with the behaviours that road users observe during interaction.

Social Appropriateness

Social appropriateness refers to how AV behaviour is interpreted by other road users during real-time interactions. It concerns whether the AV behaves in ways that people on the road recognise as appropriate. Although these judgements occur in the moment between the AV and nearby pedestrians, cyclists, or drivers, they are shaped by broader expectations within society and local communities about how road users should behave.

Several experts explained that social appropriateness depends on how traffic rules are interpreted in everyday practice. One expert noted that AVs should evaluate their actions in light of how other road users interpret legality, since rule-following can influence whether behaviour appears safe or fair (ID 1). Another expert emphasised that certain actions, such as stopping at a red light even when there is no immediate hazard, express respect for societal expectations (ID 6). These views show that compliance with traffic rules has social meaning in addition to legal meaning, and that people often use visible rule-following to judge whether behaviour aligns with shared expectations.

Experts also pointed to interpersonal qualities of behaviour. One expert stated that AVs should reflect human social values such as politeness in their driving style (ID 7). Politeness, as used by the expert, relates to driving in ways that other road users experience as considerate and cooperative. This includes avoiding abrupt or intrusive manoeuvres that could disrupt others' movement or sense of mutual accommodation.

Another expert highlighted that appropriateness also involves accounting for local norms. They explained that AVs may need to balance broad ethical principles, such as avoiding harm, with practices that are specific to local communities (ID 8). This indicates that legitimate behaviour is partly context-dependent: an action may align with expectations in one community but not in another, and recognising this variation is important for maintaining trust.

Finally, experts noted that the appearance of behaviour matters. One emphasised that AVs

should avoid actions that feel threatening to others (ID 11). This treats social appropriateness not only as a matter of rule interpretation or courtesy, but also as avoiding behaviours that could create fear or discomfort among pedestrians or other road users.

Taken together, these expert perspectives show that social appropriateness involves AV behaviour that aligns with how people expect road users to act. At the normative level, experts described socially appropriate behaviour as acting in ways that reflect the expectations of society and local communities regarding respectful and fair conduct. At the tactical level, they emphasised observable behaviours—such as courteous driving, predictable signalling, or avoiding threatening movements—that pedestrians, cyclists, and other drivers interpret in real time. This shows that social appropriateness depends on linking widely shared expectations about appropriate conduct with the moment-to-moment behavioural cues that road users rely on during interaction.

Environmental Impact

Environmental impact concerns how the AV's behaviour can contribute to or reduce environmental harm. The public and local communities generally expect AVs to demonstrate environmentally responsible driving behaviour, including reducing emissions, supporting smooth traffic flow, avoiding unnecessary fuel consumption, and acting in ways that align with community sustainability priorities.

One expert described environmental impact as a reason for AV behaviour. They recommended that AVs should minimise traffic disruption and reduce emissions, noting that smoother movement helps avoid inefficient stopping patterns and cuts fuel consumption (ID 17). The expert also linked these actions to safety, explaining that unstable flow can increase the likelihood of unsafe interactions. The justification indicates that environmental concerns are tied to how AV behaviour influences broader conditions on the road.

These expert contributions correspond to the behavioural levels used in our analysis. Environmental impact involved several behavioural levels. At the normative level, experts described environmentally responsible behaviour as acting in ways that reflect public expectations about sustainability and reduced harm. At the strategic level, they emphasised planning AV behaviour to support steady movement and minimise unnecessary energy use. These perspectives show that environmental impact depends on how the AV links broad sustainability expectations with behavioural planning that influence emissions.

Fairness

Fairness concerns how the AV should treat different people and groups without bias. Experts described this category as relating to how AV behaviour can avoid discrimination, protect those who are more vulnerable, and ensure that safety and access are not distributed unequally.

One expert stated that AVs should adapt their driving behaviour to account for the vulnerability of certain road users (ID 1). The expert explained that this is important for avoiding harm to people who face higher physical risk, such as pedestrians or cyclists. The focus was on recognising differences in exposure and adjusting behaviour accordingly.

Another expert discussed fairness in relation to infrastructure. They argued that AV-only lanes and similar designs could create inequality by restricting access or shaping mobility conditions in ways that disadvantage some groups (ID 5). The expert treated the avoidance of such outcomes as part of ensuring that socio-economic status does not determine who benefits from automated systems.

An additional expert emphasised that AVs should avoid causing harm to nearby people and animals (ID 13). Although the reason refers broadly to welfare rather than to a specific social group, it was framed as a matter of protecting those who may be vulnerable during interaction with the vehicle.

Experts also linked fairness to system design. One expert noted that AVs should not blame or penalise inattentive users and should be designed to provide fallback safety even when people make mistakes (ID 9). In this view, fairness concerns responsibility allocation and the need to ensure that design limitations do not disproportionately affect certain users.

Two experts described fairness more explicitly as a matter of equal treatment. One argued that pedestrians and occupants should be treated equally regardless of wealth or identity (ID 10). Another stated that safety standards should not vary with a user's socio-economic status (ID 12). These accounts highlight fairness as a requirement that all users receive the same protection, independent of individual characteristics.

These expert accounts align with the behavioural levels used in our analysis. Fairness emerged across multiple behavioural levels. At the normative level, experts described fairness as requiring AV behaviour that provides equal consideration to all people and avoids discriminatory outcomes. At the strategic level, they emphasised behavioural decisions shaped by system and infrastructure design, influencing who benefits from the system and how access and safety features are distributed. At the tactical level, their examples highlighted real-time behavioural adjustments—such as protecting vulnerable road users or preventing harm during interaction. Overall, these perspectives show that fairness depends on recognising differences in vulnerability and ensuring that AV behaviour does not produce outcomes that advantage or disadvantage particular users based on their social or economic position.

Efficiency

Efficiency concerns how the AV supports effective movement through traffic while aligning with the user's travel goals. Experts described this category as relating to how the AV manages speed and flow to enable timely travel without introducing unnecessary risk or disruption. They highlighted that efficiency depends both on occupant expectations for reaching destinations reliably and on design decisions that enable smooth and stable movement within traffic.

One expert explained that AVs should balance speed and responsiveness with the need to maintain efficiency throughout the trip (ID 1). The expert noted that efficiency should not come at the expense of safety, indicating that the AV must manage its pace in a way that supports steady progress without creating hazards. Another expert described efficiency in relation to user goals. They stated that AVs should consider the user's intention for the trip—such as wanting to reach a destination sooner—when selecting actions (ID 6). This perspective connects efficiency with the passenger's preferences and how the AV plans its route or driving behaviour to match them.

Table 2.3: Expert-elicited reasons, organised by reason category (rows) and behavioural levels (columns): normative, strategic, tactical, and operational. Each cell summarises what experts believed the AV should do or consider in that category. *Not mentioned* indicates no corresponding expert response for that level. Parenthetical codes indicate the human agents associated with behaviour at that level: (L) = Lawmaker/Policymaker/Regulator (RU) = Road Users, (VRU) = Vulnerable Road Users, (D) = Designer, (O) = Occupant, (GP) = General Public, (LC) = Local Community

Reason Category	Moral / Normative	Strategic	Tactical	Operational
Rule Compliance	- Follow traffic rules as societal duty and moral baseline (L) (ID 1,3,11,12)	*Not mentioned*	- Behave like a predictable "good driver" (RU) (ID 18)	- Alert and log red-light running (D) (ID 16)
Social Appropriateness	- Visible compliance signals fairness, integrity, cultural respect (RU,GP,LC) (ID 1,6,8)	*Not mentioned*	- Courteous, non-threatening driving to gain trust (RU) (ID 7,11)	*Not mentioned*
Environmental Impact	- Reduce emissions for sustainability (GP, D) (ID 17)	- Smooth traffic flow to cut fuel / stop-and-go (D) (ID 17)	*Not mentioned*	*Not mentioned*
Fairness	- Equal treatment (GP, RU) - Protect VRUs and animals (GP, RU) - Designers bear safety duty (GP, RU) (ID 9,10,13)	- Avoid exclusionary AV-only lanes (D) - Ensure safety functions for all (D) (ID 5,12)	- Adjust driving to shield vulnerable users (VRU) (ID 1)	*Not mentioned*
Efficiency	*Not mentioned*	- Balance speed with trip efficiency (D,O) - Honour faster-arrival goals (D,O) (ID 1,6)	- Avoid over-caution that disrupts flow (RU) (ID 16)	*Not mentioned*
Acceptance	*Not mentioned*	- Ensure passenger safety and comfort for acceptance (O) (ID 5)	- Drive in ways passengers find comfortable and acceptable (O) (ID 5)	*Not mentioned*
Cultural Adaptation	- Uphold legal standards despite unsafe local habits (L, LC) (ID 17)	- Adapt to region-specific traffic behaviours (D, RU) (ID 8)	- Adapt yielding / right-of-way to local norms (RU) (ID 8)	*Not mentioned*
Safety	- Adapt speed if safer (D, GP) (ID 17)	- Limit speed in pedestrian zones (RU) - Minimise manoeuvres (RU) - Plan for sensor-failure contexts (RU) (ID 4,13,15)	- Safe overtake / merge (RU) (ID 1) - Anticipatory VRU buffers(VRU) (ID 2) - Early hazard signal and brake for violator(RU) (ID 14,16) - Extra buffer when visibility blocked (RU) (ID 2)	- Detect and signal sudden obstructions early (D) (ID 14)
Interaction Management	- Transparent communication upholds fairness (D, GP) (ID 15)	- Distinguish reactive vs. goal-driven actions (D) (ID 13)	- Detect users, infer intent, and signal manoeuvres (O) (ID 3,9,15) - Predict pedestrian motion (VRU) (ID 4,7,8) - Cooperative merge / influence traffic (RU) (ID 7,18) - Decelerate to signal crossing (RU) (ID 11)	*Not mentioned*

Continued on next page

Table 2.3: Expert-elicited AV behaviour expectations (continued)

Reason Category	Moral / Normative	Strategic	Tactical	Operational
Comfort	*Not mentioned*	- Integrate safety, efficiency, comfort [D, O] (ID 7)	- Maintain comfort to avoid overrides [O] (ID 5) - Smooth merge / decel [O, RU] (ID 13) - Avoid harsh braking and needless yielding [O] (ID 9)	*Not mentioned*
Continuous Control	*Not mentioned*	*Not mentioned*	*Not mentioned*	- Maintain continuous environmental monitoring [D] (ID 14)
Control Transition	*Not mentioned*	*Not mentioned*	*Not mentioned*	- Give clear, timely takeover warning [D, O] (ID 5)
Vigilance & Readiness	*Not mentioned*	*Not mentioned*	*Not mentioned*	- Do not depend on continuous driver alertness [D, O] - Manage engagement [D, O] (ID 14)

A further expert highlighted the effect of overly cautious behaviour on traffic flow. They recommended that the AV should avoid unnecessary hesitation that disrupts surrounding traffic or leads to inefficient movement patterns (ID 16). This reflects a view that efficiency includes how the AV interacts with others and maintains stability within the broader flow of travel.

These expert views correspond to the behavioural levels used in our analysis. Efficiency appeared at the strategic and tactical behavioural levels. At the strategic level, experts described behaviours related to planning and routing that balance speed with trip efficiency and support user goals, reflecting the involvement of both designers and occupants. At the tactical level, they highlighted real-time driving behaviour that avoids overly cautious actions or unnecessary hesitation that could disrupt surrounding traffic, which road users depend on for stable flow.

Acceptance

Acceptance concerns how the AV's behaviour is experienced by passengers and by society over longer time scales. While at first glance this may appear similar to social appropriateness, the focus here is different: this category does not address how other road users interpret the AV in real-time interactions, but rather whether people feel comfortable trusting and adopting AV technology at all.

One expert explained that AVs should balance safety with long-term user comfort and acceptance (ID 6). In this view, safety remains essential, but acceptance also depends on whether the vehicle behaves in a way that passengers find comfortable and reassuring over time. The expert's description indicates that user acceptance involves both comfort considerations and a sustained sense that the AV behaves safely across repeated experience. Because acceptance develops over time, this category concerns how passengers and society evaluate the AV's behaviour beyond any single moment in traffic.

These statements fit the behavioural levels applied in our analysis. Acceptance appeared at the strategic and tactical behavioural levels. At the strategic level, experts emphasised behaviour related to balancing comfort and safety over time so that people can rely on the system. At the tactical level, they highlighted specific driving behaviours that influence passenger comfort and confidence during use.

Cultural Adaptation

Cultural adaptation concerns how the AV should account for location-specific behavioural expectations and informal practices in different traffic environments. Experts described this category as relating to the way norms vary across places and how these variations shape what local road users and local community expect from an AV.

One expert explained that AVs should adapt their behaviour to reflect local cultural norms and road conventions so that their actions match what people in that environment consider appropriate (ID 8). This includes modifying responses to align with expectations that differ across locations, such as how pedestrians or cyclists typically behave in that region.

The same expert illustrated how local expectations influence right-of-way decisions by comparing cyclist behaviour in the Netherlands and the United States. They explained that if be-

haviour were guided only by a rule such as “do not harm”, a Dutch cyclist would stop whenever a pedestrian might cross their path. In practice, cyclists in the Netherlands usually continue unless the pedestrian clearly commits to entering the road. In California, however, continuing in this way could be viewed as improper or even immoral. The expert used this to explain that AVs should adjust their responses based on location-specific expectations about yielding and movement patterns (ID 8).

Another expert highlighted situations where local driver behaviour may be inconsistent with formal rules. For example, they noted that an AV should obey traffic laws even when local drivers act unpredictably in features such as roundabouts (ID 17). The expert emphasised that the AV should not reproduce informal or unsafe habits, even when these habits are common, but should still recognise them in order to navigate reliably.

In summary, experts discussed cultural adaptation in relation to different behavioural levels. At the normative level, experts described expectations grounded in local values about what counts as appropriate behaviour in a given place. At the strategic level, they referred to the need for AVs to adjust decisions to match location-specific road conventions, including differences in right-of-way expectations. At the tactical level, they discussed how the AV should respond to behaviours that occur in real time, such as informal practices at roundabouts or crossings, while still respecting legal requirements. These observations show that AVs are required to understand local practices while maintaining behaviours that remain consistent with legal and ethical expectations when local norms diverge from them.

Safety

Safety concerns how the AV avoids unsafe situations and reduces the likelihood of harm for both the people directly interacting with the vehicle and the wider public who depend on safe road systems. Experts described safety as relating to how the AV anticipates threats, manages uncertainty, and responds in ways that limit the possibility of collisions and support public expectations of safety.

One expert stated that AVs should execute overtaking and merging manoeuvres in ways that reduce crash likelihood (ID 1). The expert explained that safe positioning during interaction is necessary for limiting immediate risk on the road.

Another expert described how AVs should act when the presence of vulnerable road users is uncertain. They noted that the AV should perform anticipatory manoeuvres and leave buffer space when visibility is limited, so that unexpected encounters do not lead to unsafe situations (ID 2). This reasoning emphasised caution when information about the environment is incomplete.

Experts also referred to planning behaviour around pedestrians. One expert explained that manoeuvres should be planned to limit speed and create enough space for pedestrians to pass safely (ID 4). Similar points were made about planning based on the status of the AV and the position of surrounding road users so that manoeuvres are informed by the conditions of the environment (ID 13).

Other contributions focused on how the AV should detect and communicate changes in the environment. One expert stated that early detection of hazards helps the AV mitigate risk before

unsafe situations develop (ID 14). Another expert described how unnecessary manoeuvres can introduce additional risk, and suggested limiting such behaviour when safe progress can still be maintained (ID 15). They also noted that when sensor reliability is compromised, the AV should assess the situation in full before responding (ID 15). These points emphasised that risk management includes adapting to changing conditions.

Some experts highlighted behaviours relevant to exceptional or unpredictable scenarios. One expert noted that the AV should detect red-light violations by others and brake when needed to avoid collision (ID 16). The same expert explained that, in uncertain situations, the AV should prioritise protecting its occupants (ID 16). Another expert stated that adapting to realistic traffic speeds, rather than relying only on strict legal limits, can support safer movement when surrounding flow differs from posted rules (ID 17).

These observations relate directly to the behavioural levels defined in our analysis. Safety appeared across several behavioural levels. At the normative level, experts referred to expectations from the wider public that the AV should prioritise protecting both occupants and other road users when risk is present. At the strategic level, they described planning manoeuvres that account for pedestrian movement, speed limits, sensor performance, and road-user positioning, reflecting the design decisions that shape safe movement. At the tactical level, their examples focused on real-time adjustments such as leaving buffer space, adapting to occlusion, responding to red-light violators, and braking when hazards emerge. At the operational level, experts highlighted behaviours such as detecting sudden obstructions. Overall, these perspectives show that safety depends on how the AV links protective priorities with planned manoeuvres and immediate responses to unfolding events.

Interaction Management

Interaction management refers to how the AV interprets the actions of people outside the vehicle and makes its own behaviour understandable to them. Experts consistently described this category as concerned with how the AV interprets the behaviour of pedestrians, cyclists, and other drivers, and how it makes its own behaviour understandable to them.

Several experts emphasised that the AV should detect relevant road users and infer their intentions so that it can respond in a way that avoids unsafe interaction (ID 3). This includes understanding whether others are slowing, crossing, or changing direction. The expert framed this as a necessary part of responding appropriately to surrounding behaviour.

Experts also described the importance of communication during interaction. One expert explained that the AV should convey its intended manoeuvres clearly to ensure that other road users can anticipate what it will do (ID 3). Another noted that the use of visible cues, such as cinematic indicators or clear deceleration, helps pedestrians interpret the AV's movement and judge whether they can proceed (ID 9). A similar point was made regarding hazard signals, where one expert stated that clear warnings help prevent misinterpretation during abnormal situations (ID 15).

Anticipating the movement of others was another common theme. One expert highlighted that the AV should predict pedestrian motion at both marked and unmarked crossings to prevent unsafe encounters (ID 4). Other experts provided similar examples involving the prediction of vehicle, cyclist, or scooter trajectories in more complex settings such as intersections (ID 8, ID

7). These points treat anticipation as a central part of managing shared road space.

Experts also referred to how the AV's behaviour affects the actions of others. One expert explained that the AV should influence the behaviour of surrounding road users by using cooperative or predictable manoeuvres (ID 7). Another described merging behaviour as an example, noting that the AV should cooperate with other vehicles during such manoeuvres so that its behaviour aligns with what others expect (ID 18). These examples show how interaction management involves shaping not only the AV's responses but also how others adjust their behaviour.

These perspectives reflect the behavioural levels that structure our analysis. Interaction management appeared across several behavioural levels. At the normative level, experts framed clear and transparent signalling as a behavioural responsibility that supports fairness and public trust in shared road environments. At the strategic level, they described behavioural decisions about how the AV positions itself, prepares manoeuvres, and distinguishes between reactive and goal-directed actions so that its behaviour fits the broader flow of traffic. At the tactical level, their examples highlighted real-time behaviours such as detecting nearby road users, interpreting their intentions, predicting their movement, and using signalling or cooperative manoeuvres to make the AV's actions understandable. Overall, these views show that interaction management depends on how the AV interprets others' actions, communicates its own intentions, and coordinates behaviour within dynamic shared road space.

Comfort

Comfort concerns how passengers perceive the AV's driving behaviour and how this perception shapes both their immediate reactions and their longer-term experience of automated travel. Although this category may appear similar to interaction management, the focus here is different: comfort concerns the effects of the AV's movement on people inside the vehicle, whereas interaction management addresses how people outside the vehicle interpret and respond to what the AV does.

These perspectives reflect the behavioural levels that structure our analysis. Several experts described comfort as closely connected to safety because discomfort can prompt unnecessary intervention. One expert explained that the AV should maintain passenger comfort to prevent overrides caused by uncertainty or confusion (ID 5). This perspective treats comfort as part of maintaining a stable relationship between the passenger and the vehicle, since discomfort may lead to actions that interfere with automated control.

Another expert noted that the AV should drive in a way that feels comfortable to passengers and intuitive to nearby road users (ID 13). In this account, comfort supports acceptance by ensuring that the vehicle's motion aligns with what passengers expect and what other road users can understand.

Experts also mentioned that comfort should be considered when planning actions. One expert described comfort as a factor the AV should address alongside safety and efficiency when determining how to act (ID 7). This reflects the idea that comfort is part of how the AV should structure its behaviour over time, not only in immediate responses.

Two examples from another expert concerned specific driving practices that influence com-

fort. Avoiding harsh braking was described as important for passengers on board (ID 9). The same expert noted that the AV should avoid yielding when doing so would create unnecessary disruption for the occupants (ID 9). These examples show how comfort appears in particular manoeuvres.

Taken together, these contributions describe comfort as aspects of AV behaviour that shape how passengers interpret and respond to automated driving. Comfort influences whether passengers feel secure and whether they choose to intervene. It also affects how understandable the AV's movement appears to people who interact with the vehicle.

These statements fit the behavioural levels applied in our analysis. Comfort mapped onto strategic and tactical behavioural levels. At the strategic level, they referred to comfort as something the AV should incorporate when planning how to act over longer stretches of driving. At the tactical level, their examples focused on how specific manoeuvres—such as braking or yielding—affect the passenger's immediate experience. This shows that comfort operates at several levels of behaviour, shaping how passengers respond both in individual moments and across sustained interactions with the AV.

Continuous Control

Continuous control concerns how the AV sustains awareness of its surroundings and remains responsive during driving so that its behaviour is predictable and safe for other road users, particularly those who are vulnerable. Experts described this category as relating to the AV's ability to monitor traffic conditions continuously rather than relying only on discrete updates or isolated events, which depends on design choices that enable consistent awareness throughout the trip.

One expert stated that the AV should maintain attention to changes in the traffic environment even during routine driving (ID 14). They explained that doing so supports ongoing awareness and allows the AV to respond when conditions shift. The reason emphasised that continuous monitoring is necessary for timely and appropriate adjustment to what happens around the vehicle.

The statement aligns with the operational task from behavioural levels. Experts described the need for ongoing monitoring and real-time responsiveness as traffic conditions change. This suggests that continuous control, as described by experts, centres on maintaining persistent situational awareness and the capacity for immediate behavioural adjustment in response to evolving traffic conditions.

Vigilance and Readiness

Vigilance and readiness concerns how the AV manages attention and responsibility in situations where control is shared between the vehicle and the human driver. Experts described this category as relating to whether the AV should depend on the driver, who may not remain continuously alert during extended periods of automation, and to the role of system designers in determining how responsibility is allocated when attention declines.

One expert stated that the AV should avoid relying on driver alertness in long-term shared-

control situations (ID 14). They explained that drivers often become inattentive when automation manages the majority of the driving task and that expecting the driver to remain vigilant under these conditions is unsafe. The reason emphasised the need for the AV to handle responsibility directly when prolonged automation reduces the likelihood of sustained human attention.

Vigilance and readiness were discussed only at the operational level, where experts described the need for the system to always remain alert, avoid depending on the driver, and maintain driver engagement. This suggests that vigilance and readiness, as described by experts, centre on the dual tasks of ensuring the driver remains engaged and prepared for operational demands, while the AV itself remains continuously attentive to the surrounding environment.

Control Transition

Control transition concerns how responsibility shifts between the AV and the human driver, and how this transition is supported by the system's design. Experts discussed this category in relation to how the AV prepares the driver to resume manual control safely, emphasising the role of designers in determining how handover is communicated and managed.

One expert stated that the AV should provide sufficient warning before the driver takes back control (ID 5). They explained that this is needed to prevent confusion and to ensure that the driver can safely resume the task. The expert emphasised that a clear and timely transition reduces the likelihood of unsafe responses when authority over the vehicle changes.

This explanation only aligns with the operational task of behavioural levels. The expert highlighted the need for clear and timely cues that allow the driver to safely resume control. This shows that control transition centres on how the AV provides real-time support during the moment when responsibility shifts to the driver.

2.3.3 Discussion

To address Objective 1, we aimed to identify and structure the category of human reasons that should inform AV manoeuvre planning. Our study contributes to addressing the **practical gap** by offering a layered mapping of thirteen reason categories, organised across moral, strategic, tactical, and operational levels, and explicitly linked to the roles of relevant human agents. This structure organises the expert-derived human reasons that influence AV manoeuvre planning into a layered form of guidance on what considerations should inform AV decision-making, supporting future research and system design aligned with the tracking condition of Meaningful Human Control (MHC).

This section also responds to the methodological gap identified in the introduction by demonstrating how expert interviews and directed content analysis can reveal the structure of human reasons relevant to AV behaviour. Our approach systematically captured how experts from diverse domains interpret what matters in AV behaviour, allowing reason categories to emerge inductively while using the Meaningful Human Control framework to position them across behavioural levels. Table 2.3 illustrates how these reasons connect to AV behaviour across different levels, from normative expectations to operational execution, and identifies the human agents affected by those behaviours. This mapping provides a bridge between the reasons experts expressed and the types of behaviour designers may need to support in AV systems.

Multiple overlapping reasons in a single manoeuvre. A central insight from our findings is that AV manoeuvre planning rarely relies on a single type of reason. Instead, a single manoeuvre often engages multiple overlapping reasons that reflect different layers of ethical and practical concern. For example, Expert ID 17 explained that choosing not to follow traffic rules strictly in some situations can be justified for more than one reason at the same time, such as improving safety and reducing environmental impact. Additionally, reason categories themselves are not confined to one behavioural layer. Rule compliance, for instance, spans normative expectations (e.g., respecting laws), tactical execution (e.g., behaving like a predictable driver), and operational functionality (e.g., logging red-light violations). This layered nature of reasons suggests that manoeuvre planning systems must support multi-reason, multi-level responsiveness, rather than relying on rule-based execution alone.

Human proximity and agent roles. Our analysis of Figure 2.1 indicates a relationship between reason type and the proximity of the human agents referenced by participants. Normative reasons were typically associated with more socially and institutionally distant agents—such as policymakers, the general public, and local communities—who shape and enforce broader ethical standards. In contrast, strategic, tactical, and operational reasons were more closely connected to agents physically proximate to AV operation and who directly interact with or design AV behaviour, such as vehicle occupants and system designers. Notably, road users and vulnerable road users appear across the full range from normative to operational levels, suggesting that participants viewed them as important agents to consider throughout all layers of reasoning. This supports and extends the proximity-based model introduced by Mecacci & Santoni de Sio (2020) and elaborated by Calvert & Mecacci (2020), which posits that meaningful human control depends on responsiveness to human reasons distributed across different layers of reasoning. Our findings provide empirical evidence for this framework and offer a structured account of how agent proximity to the AV and the type of reasons are interconnected.

Variation in behavioural level depending on task interpretation. We also found that the behavioural level at which a reason is situated can vary depending on how the AV task is interpreted. For example, time efficiency is often treated as a strategic concern, but depending on the situation, it may also appear at tactical or even operational levels. A strategic interpretation might involve planning the most efficient route, while a tactical one may involve decisions such as overtaking or avoiding hesitation that disrupts flow. Despite being motivated by the same underlying reason of efficiency, these interpretations correspond to different layers of action. This highlights the importance of distinguishing between the justification for a behaviour and the specific behavioural level at which it is operationalised.

Reason variations across levels of automation. We also observed that the relevance of certain reasons shifts with the AV's level of automation. For instance, considerations such as control transition and vigilance were particularly prominent for lower levels of automation (L2/L3), where human fallback is still required. At higher levels (L4/L5), these concerns recede, and the focus shifts towards trade-offs among values such as fairness, efficiency, and comfort—especially when AVs operate without direct human supervision. Our framework accommodates these shifts by revealing which reasons—and which agents—are most relevant at each automation stage and behavioural level.

Interpretative flexibility in core categories. A further nuance in our data involves the diverse interpretations of rule compliance. While several experts (e.g., IDs1, 3, 11, 12) framed rule-following as a strict moral baseline, others (e.g., ID17) viewed traffic laws as flexible guidelines to be overridden when necessary to ensure fairness or safety. This reflects the contextual and scenario-sensitive nature of AV ethics: manoeuvre planning must not only track rules but also balance them against competing values such as social appropriateness and safety.

Positioning within existing literature. Our framework also offers a way to contextualise and relate existing efforts to define what AVs should consider when making driving decisions. Prior work has provided focused contributions on specific types of consideration: for example, UN-ECE (2023); Olleja et al. (2025) define legal and behavioural benchmarks for competent driving; Geisslinger et al. (2021) formalise ethical principles for risk-sensitive planning; Schwarting et al. (2018) model social value orientation for cooperative behaviour; and Thornton et al. (2018) apply value-sensitive design to embed stakeholder values in AV system logic. While these approaches differ in their aims and methodologies, our framework does not attempt to replace or rank them. Rather, it provides a layered structure through which these contributions can be situated—by connecting the types of reasons they represent (e.g., legality, fairness, efficiency, comfort) to specific levels of AV behaviour (e.g., moral/normative, strategic, tactical, operational). In this sense, our empirically derived categorisation can serve as a common referential model—one that helps clarify how diverse AV design goals and values interact across system layers and in relation to different human agents.

Practical design guidance and limitations. Beyond theoretical insights, our structured mapping offers practical guidance for AV developers and policy designers. For example, developers working on L2/L3 vehicles may use our findings to prioritise clear and timely takeover cues, while those building L4/L5 systems may focus on fairness–efficiency trade-offs and behaviour intelligibility in mixed traffic. This structured mapping of thirteen reason categories across behavioural levels and agent roles can help translate ethical expectations into system-level specifications—by clarifying what kinds of human concerns should be considered, at which layer of system behaviour, and by whom. Furthermore, as the questions did not solely focus on ethically challenging situations, the identified reasons are applicable not only in edge cases but also in routine and general AV driving scenarios where aligning AV behaviour with human reasons is essential.

Nonetheless, our approach has limitations. First, the scope of this study is restricted to manoeuvre-level decision-making, rather than broader systemic influences such as infrastructure, corporate strategy, or legal frameworks. Second, interpretations of the identified reasons are likely influenced by cultural and regional contexts, which may affect the generalisability of the findings. Our expert pool was primarily composed of individuals from Western countries, particularly the Netherlands, the United States, and the United Kingdom. As such, our findings should be understood as reflecting predominantly Western ethical and social norms, and not assumed to be universally applicable. Future research is needed to explore how these categories of reasons manifest in other cultural environments.

Third, we emphasise that the 13-category taxonomy is not intended as exhaustive. It reflects insights from a specific group of experts, and future research should investigate how cultural, regional, or stakeholder diversity may yield additional or context-dependent reason types. In particular, cells marked “Not mentioned” in Table 2.3 do not imply that no relevant reason-

ing exists at that level. Theoretical frameworks suggest that any reason could, in principle, be interpreted across multiple control layers. However, because our study is empirical in nature, the absence of content in certain cells reflects the limits of what was raised by experts, not a conceptual impossibility. Future research could further investigate these gaps through targeted questioning or normative modelling. Future work may also extend this approach by incorporating a broader and more diverse stakeholder base—such as regulators, insurers, and urban planners—or by developing prioritisation models to resolve conflicts among overlapping reasons.

Finally, while every effort was made to ensure conceptual clarity, we acknowledge that some reason categories, such as *social appropriateness* and *acceptance*, may overlap in practice. This reflects the interpretive nature of qualitative analysis. However, these overlaps do not affect the core findings regarding how experts prioritise reasons in ethically ambiguous situations. Consensus coding and iterative refinement were used to mitigate interpretive bias, and we encourage future research to further validate and refine the categorisation scheme using participatory or quantitative methods.

Summary of contributions. In summary, addressing Objective 1, we identified thirteen categories of human reasons relevant to AV manoeuvre planning and examined how these reasons informed expert expectations about what the AV should do across different behavioural layers, as well as which human agents were affected. Our findings highlight that even routine AV decisions can involve ethically sensitive considerations, and that manoeuvre planning must account for multiple, sometimes conflicting, human expectations. By integrating these findings with the MHC framework, our study offers a pathway towards AV systems that behave in ways aligned with human reasons.

2.4 Reason-Based Decision Principles for Ethically Challenging AV Scenarios: Cyclist Overtaking as a Case Study

2.4.1 Methods

To investigate reason-based decision principles in ethically challenging cyclist overtaking situations, we analysed expert responses to Questions 5–11 of the semi-structured interview protocol administered to the same participants described in Subsection 2.3.1. This section outlines the scenario design, question format, and the analytical approaches used to examine both implicit and explicit prioritisation of reasons.

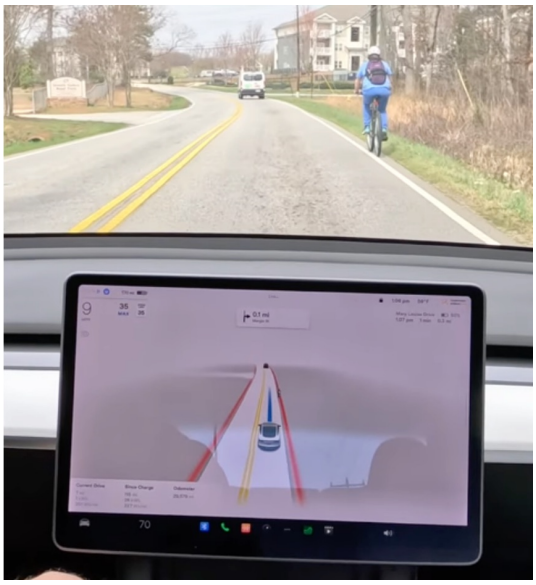
2.4.2 Scenario and Questionnaire Design

To explore the reason-based principles for AV decision-making in ethically challenging traffic situations, we incorporated a specific overtaking scenario in our interview protocol. This scenario involved an AV travelling behind a slow-moving cyclist on a two-way road marked with a double yellow line. This road marking typically prohibits overtaking, thereby introducing a regulatory constraint that renders the situation ethically ambiguous: the AV could either remain

behind the cyclist at a reduced speed or initiate an overtaking manoeuvre by crossing the double yellow line.

After responding to Questions 2–4, which explored their reasoning about what the AV should consider when planning its manoeuvre, experts were shown a short video clip depicting the overtaking scenario (see Table 2.4). They then answered a series of structured follow-up questions (5-9) that asked them to predict how the AV and other traffic participants would behave, explain the reasons for these actions, identify additional influencing factors, and describe possible conflicts between different intentions. Finally, in questions 10-11, experts ranked the intentions of three predefined stakeholders in the scenario (the AV passenger, the cyclist, and the road policymaker), where "intentions" were operationalised as proxies for underlying reasons, and explained the reasoning behind their rankings. These questions were designed to elicit interpretations of AV decision-making, the factors influencing it, and how experts implicitly and explicitly prioritised competing reasons, without explicitly prompting normative judgments about what the AV should do.

Table 2.4: Interview Questions Relevant to Objective 2

Question number	**Question**
Please watch the video below and read its description  **Video description** A passenger uses an automated vehicle (AV) for a morning commute to the office. The passenger has an important meeting and must arrive on time. If the vehicle maintains the current speed, the passenger can reach the office on time in 20 minutes. The AV is on a road with solid double yellow lines, which prohibit vehicles from crossing in both directions due to safety reasons. During the trip, the AV approaches a cyclist traveling at half of the speed of the AV. There is no safe passing zone visible from the vehicle; however, the opposite lane is currently empty.	
Q5	If the video continues, what do you believe all traffic participants will do?

Continued on next page

Table 2.4 – Continued from previous page

<table>
<tr><th>Question number</th><th>Question</th></tr>
<tr><td>Q6</td><td>What are the reasons for the [traffic participants mentioned by the experts] performing the [actions the experts mentioned]?</td></tr>
<tr><td>Q7</td><td>Besides the [traffic participants that are mentioned by the experts], can you think any other factors that might influence the traffic participants' decisions?</td></tr>
<tr><td>Q8</td><td>What do you think the reasons are for the [other factors that are mentioned by the experts]?</td></tr>
<tr><td>Q9</td><td>Can you think of any situations where the intentions of the [traffic participants / other factors the experts mentioned] might conflict? Please share any examples you can think of, and let me know when these conflicts may typically occur.</td></tr>
<tr><td colspan="2">Recall the scene from the previous video.
There are three different people, each with their own intentions:
• The automated vehicle (AV) passenger wants to pass the cyclist to get to the office on time.
• The cyclist wants a safe distance from the AV for safety concerns.
• The road policymaker wants both AV and cyclist to use their designated lanes, marked by solid yellow lines, for everyone's safety.
Keep this in mind as you answer the rest of the questions.</td></tr>
<tr><td>Q10</td><td>From your perspective, whose intentions should be given the most importance? Please answer this question by ranking the individuals below, with '1' indicating the highest rank.
<table>
<tr><th></th><th>1</th><th>2</th><th>3</th></tr>
<tr><td>AV passenger</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Cyclist</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Road policymakers</td><td>◯</td><td>◯</td><td>◯</td></tr>
</table></td></tr>
<tr><td>Q11</td><td>Could you please explain the reasons behind the rank you provided in your previous answer?</td></tr>
</table>

2.4.3 Expert Reasoning in the Cyclist Overtaking Scenario

AV Behaviour Justifications

To understand how experts justify their predictions about AV behaviour in overtaking scenarios, we used *Qualitative Content Analysis (QCA)* (Schreier, 2012) to analyse their responses to Questions 5–9. These questions asked experts to predict what the AV would do in a specific overtaking scenario and to justify their predictions. Following Schreier's structured approach, we developed a mixed-category coding strategy that incorporated both:

- **Concept-driven categories**, used to group expert responses by predicted AV behaviour: either overtaking the cyclist or not overtaking the cyclist.
- **Data-driven categories**, used to inductively identify the specific justifications experts provided for their predictions.

The first stage of the analysis involved classifying each expert response according to the predicted action (i.e., follow vs overtake). Within each group, expert justifications were then collected and analysed to identify recurring patterns of reasoning. These inductively derived justifications were subsequently mapped to the predefined set of thirteen reason types originally developed from the open-ended responses to Questions 2–4 (see Section 2.3.2).

As part of the interpretation phase, we conducted a qualitative synthesis of the coded data to explore patterns in how experts linked specific reasons to predicted AV behaviours. This involved examining which reasons frequently co-occurred, how the same reasons were interpreted differently across contexts, and instances where multiple reasons appeared within a single justification. This helped us understand how reasons relate to each other or are prioritised. The synthesis was guided by principles of thematic pattern analysis (Braun & Clarke, 2006) and constructivist grounded theory (Charmaz, 2014), with analytical interpretations discussed collaboratively between authors to enhance transparency and reduce individual bias.

To analyse reason prioritisation, we examined reason prioritisation in two complementary ways. The first approach captured how priorities emerged naturally within participants' open-ended reasoning (implicit prioritisation). The second approach asked participants to directly state their preferences in a structured ranking task (explicit prioritisation). Analysing both allowed us to compare context-driven unprompted prioritisation with deliberate stated preferences. Together, these perspectives gave us a fuller understanding of how experts weigh competing reasons.

Implicit Reason Prioritisation

Implicit analysis allowed us to capture how prioritisation emerged naturally in participants' reasoning, without being influenced by a predefined ranking task or fixed response options. This approach provided insight into how trade-offs were navigated in context and how multiple considerations interacted within the flow of open-ended discussion.

Building on the qualitative synthesis described in Section 2.4.3, we conducted an analysis of implicit prioritisation patterns evident in participants' responses. Although the questionnaire did not directly ask experts to rank reasons, the overtaking scenario was intentionally designed to present conflicting reasons within a single context, particularly through the video depiction. This design allowed us to observe how participants navigated trade-offs between different considerations when explaining their expected or preferred AV behaviour.

To systematically analyse this *implicit prioritisation*, we examined how participants justified their decision for the AV to either follow or overtake the cyclist. We focused on statements where reasons were weighed against each other in explaining that decision. For example, some participants accepted a rule violation (crossing the double yellow line) because it would reduce cyclist discomfort. Others rejected an overtake because rule compliance outweighed the driver's travel efficiency. Such comparisons allowed us to infer the relative importance of reasons. In the first example, cyclist comfort was treated as a higher priority than strict rule compliance, whereas in the second example, rule compliance was treated as a higher priority than the driver's travel efficiency. From these comparisons, we identified which reasons were positioned as primary and which were conditional or secondary, even without an explicit ranking task.

Explicit Prioritisation

Explicit ranking provided a direct statement of participants' preferences, enabling systematic comparison across individuals and alignment checks with the priorities inferred from the implicit analysis. This approach also allowed us to quantify the relative importance assigned to each stakeholder's reason.

To complement the implicit analysis in Section 2.4.3, we included a structured ranking task in Questions 10-11 to explicitly elicit prioritisation preference. Experts were asked to rank the intentions (used as proxies for broader reasons, but phrased this way for participant clarity) of three human agents involved in the overtaking scenario: the AV passenger (who wants to reach the office on time), the cyclist (who wants a safe buffer zone), and the road policymaker (who designed the double yellow lines for public safety).

Experts were provided both a numerical ranking and a free-text explanation of their choices. This approach enabled us to collect both quantitative and qualitative data on how experts explicitly prioritised different stakeholder reasons. Rankings were aggregated to identify overall patterns, and justifications were thematically analysed based on the order of rank to uncover the themes guiding these decisions.

2.4.4 Results

This section presents the findings from the expert interviews. Based on their responses to Question 5-9, most experts interpreted the AV as the sole traffic participant whose actions were being evaluated. Two primary behaviours that the AV might adopt in the given scenario were identified: (1) *following the cyclist*, and (2) *overtaking the cyclist*.

In addition, some experts distinguished between what the AV is likely to do (**expected action**) and what the AV ought to do (**preferred action**). To maintain clarity, this distinction will be upheld throughout the remainder of the paper: expected action refers to what the AV is predicted to do, whereas preferred action refers to what the AV should do from the expert's normative perspective.

Figure 2.2 summarises the reasons experts provided for predicting whether the AV **will** or **should** follow or overtake the cyclist. Thirteen distinct reasons are presented, grouped by action type. Blue shading represents predicted ("will") actions, while lighter red shading represents preferred ("should") actions.

In general, experts cited a more limited set of reasons for why the AV will follow the cyclist, most commonly grounded in *rule compliance* and *safety*. By contrast, overtaking was associated with a broader range of justifications, including *efficiency*, *comfort*, and *interaction management*, alongside the aforementioned safety and legal concerns.

This pattern suggests that *following the cyclist* is predominantly justified by a narrow range of risk-averse or rule-based considerations, whereas *overtaking* is viewed as a more complex decision that draws upon a wider set of overlapping reasons. We then go beyond listing reasons to examine how experts weighted competing reasons in their explanations (implicit prioritisation) and how they explicitly ranked stakeholder reasons in a structured task (explicit prioritisation). Together, these analyses provide a fuller picture of not only what reasons experts described, but

also how they balanced them when reasoning about AV behaviour.

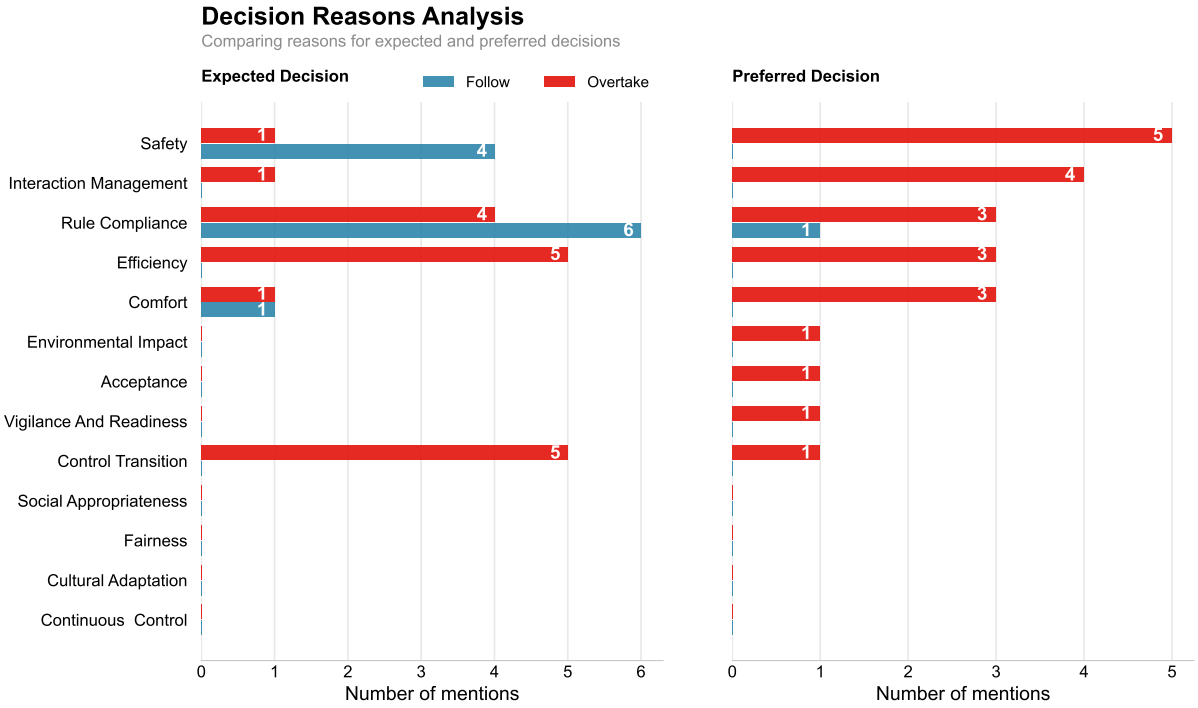


Figure 2.2: Experts' reasons for predicting whether the AV will or should follow (a) or overtake (b) the cyclist. Thirteen reasons are listed in the first column, expert IDs in the second, and total counts in the third. Blue indicates predicted behavior (will), red indicates preferred behavior (should).

Reasons for the AV Will Follow the Cyclist

From Figure 2.2, panel (a), the most frequently cited reasons for why the AV is expected to follow the cyclist were *rule compliance* and *safety*. Several experts (e.g., ID2, ID12) emphasised that the AV is programmed to obey traffic laws, such as not crossing a double yellow line, making overtaking legally impermissible. Others (e.g., ID3, ID11) pointed to the importance of safety, arguing that the AV would stay behind the cyclist to avoid potential collisions or unsafe manoeuvres. In many cases, both legal and safety considerations were closely intertwined in the experts' reasoning. One expert (ID2) also mentioned driver discomfort as a potential outcome, noting that while the AV is obligated to follow the cyclist, this may result in frustration or discomfort for the human passenger. However, this was framed not as a reason for the AV's behaviour, but rather as a consequence of its strict adherence to rules and safety protocols.

Reasons for the AV Should Follow the Cyclist

Notably, only one expert (ID12) explicitly stated that the AV *should* follow the cyclist, as indicated by the lighter red shading in panel (b). This expert argued that, in addition to the fact that current AVs are programmed to follow traffic rules, they *should* continue to be programmed in accordance with those rules. Interestingly, this expert was the only one with a background in the legal aspects of AVs.

Reasons for the AV Will Overtake the Cyclist

In panel (a) of the figure, the experts cited a variety of reasons for why the AV is expected to overtake the cyclist. The most frequently mentioned reasons included *efficiency*, the possibility of *control transition*, and *rule compliance*. *Safety* and *comfort* were discussed less frequently but still featured in some responses.

Several experts emphasised time efficiency as a key factor. One expert (ID15) stated that the driver would likely take over control and overtake the cyclist in order to arrive at work on time. Another expert (ID10) also mentioned the urgency of reaching a destination as a reason for manual takeover. A third expert (ID7) explained that while the AV might initially follow the cyclist, the driver would likely take over if delays occurred, particularly because they wouldn't want to be late. Similarly, an expert (ID13) noted that overtaking would allow the AV to maintain a more efficient speed, and another (ID14) emphasised that roads are designed for higher speeds than cyclists typically travel, making it uncomfortable for drivers to go significantly slower than expected.

Rule compliance was discussed in relation to whether overtaking is legally permissible. One expert (ID13) stated that in the UK, it is legal to overtake a cyclist if they are travelling below 10 mph, and therefore expected the AV to do so. Another expert (ID6) viewed the AV's failure to overtake when the opposite lane is empty as irrational, suggesting that such behaviour would seem "stupid" to human drivers. An expert (ID7) commented that while drivers are aware of traffic rules, they often weigh the risk of breaking those rules against the need to stay on schedule. One expert (ID15) believed that the AV would not overtake because it is programmed to strictly follow traffic rules, but that the driver would override this behaviour when time becomes a priority.

The possibility of *control transition* was also commonly mentioned. Multiple experts (IDs 6, 7, 10, 14, 15) described scenarios in which the human driver would take over control to overtake the cyclist. Some (e.g., ID7, ID10) mentioned this as a temporary takeover, while others (e.g., ID14, ID15) connected it to the frustration caused by prolonged low-speed travel. One expert (ID6) stressed that such a manoeuvre would not compromise safety and therefore believed the driver would proceed with overtaking.

A few experts raised *social and comfort*-related concerns. One expert (ID13) emphasised the importance of not frustrating other road users who may be following the AV. Another expert (ID14) pointed out that drivers are not accustomed to travelling so slowly, especially on roads designed for higher speeds, and would likely feel discomfort in such situations—prompting a takeover.

Although less frequently mentioned, *safety* still featured in the discussion. One expert (ID6) stated that overtaking would be acceptable if the opposite lane were empty, implying that safety would not be compromised. Another (ID10) said the driver would overtake only if it could be done without putting anyone at risk, suggesting that manual takeovers are still constrained by a concern for *harm avoidance*.

Reasons for the AV Should Overtake the Cyclist

Several experts argued that the AV *should* overtake the cyclist, even if it involves crossing a double yellow line, because failing to do so could cause confusion, frustration, and even safety issues. A common theme among the responses was that not overtaking could disrupt traffic flow, potentially leading to cascading negative effects—such as increased risk (ID1, ID17), discomfort (ID5, ID11), higher emissions (ID17), and reduced acceptance (ID11). These disruptions were framed not only as inefficiencies but also as factors that could compromise safety and overall system performance.

For some experts, *safety* did not necessarily align with strict rule-following. Instead, safety was interpreted contextually, such as maintaining adequate distance from the cyclist (ID8), supporting a steady traffic flow (ID16, ID17), or reducing cognitive workload for drivers (ID5). This suggests that pragmatic, situational decisions—even if technically in violation of traffic regulations—can still serve safety-related objectives.

Experts also emphasised that a single justification was rarely sufficient; rather, multiple factors often combined within a single rationale. For example, an expert might state that the AV should overtake because it is both *safer* and *more comfortable* (ID5), or because it improves *traffic flow* and aligns with *human driving behaviour* (ID17). These reasons were presented as equally important, rather than hierarchically ordered.

Regarding *rule compliance*, experts acknowledged that overtaking may violate traffic rules (ID8, ID11, ID16). However, they generally agreed that strict rule adherence should not override other practical concerns such as *safety* and *traffic efficiency*. In this context, rule violations were often framed as acceptable when they led to better outcomes for all road users.

In addition, *interaction management* and *comfort* played significant roles in shaping expert expectations. Some experts noted that cyclists may feel uncomfortable or stressed when a vehicle follows too closely without overtaking (ID5, ID11, ID17), and that passengers or drivers may become frustrated by overly cautious AV behaviour (ID1, ID11). These concerns link comfort with public trust and acceptance, suggesting that AVs should behave in ways that are intelligible and relatable to human road users.

Finally, some experts (e.g., ID5) stated that in such situations, they would personally choose to overtake the cyclist by taking control of the vehicle. This reinforces the view that manual takeover may remain a practical necessity when AVs are constrained by rules that fail to account for situational flexibility.

Implicit Prioritisation Patterns

Analysis of participants' explanations revealed consistent patterns in how reasons were prioritised when predicting or prescribing AV behaviour in the overtaking scenario. Across interviews, safety consistently emerged as the primary, non-negotiable consideration. Even when participants supported overtaking, they emphasised that it should only occur if safety could be maintained. For example, some stated that overtaking was only acceptable if the opposite lane was clear (ID6) or if there was adequate distance from the cyclist (ID8, ID10).

Other reasons, such as efficiency, comfort, and interaction management, were frequently mentioned, but typically in conjunction with safety. These were often framed as benefits that

could be achieved only if the manoeuvre met safety conditions. For instance, participants linked overtaking to maintaining steady traffic flow and reducing emissions (ID17), preventing frustration for following drivers (ID1, ID13), and relieving stress for cyclists (ID5, ID11, ID17).

Rule compliance was generally treated as a conditional obligation. Some participants cited it as a reason for the AV not to overtake, noting that the system would be programmed to follow the double yellow line rule (ID2, ID3, ID12). Others described it as the very reason a manual takeover might occur, since human drivers could choose to overtake when the AV, bound by traffic laws, would not (ID7, ID10, ID15). Many argued that crossing the line could still be justified when safety and traffic flow benefits outweighed strict adherence (ID6, ID7, ID13, ID16).

A smaller set of reasons, including environmental impact, driver workload reduction, and maintaining steady traffic flow, appeared less frequently and were often context-specific (ID5, ID14, ID17).

Overall, participants tended to treat safety as a primary consideration; efficiency, comfort, and human-interaction concerns were usually framed in relation to safety; and rule compliance was often conditional—sometimes cited as the reason the AV would not overtake, leading to manual takeover by the driver, and at other times set aside when safety or traffic flow benefits outweighed strict adherence (Figure 2.2).

Explicit Prioritisation Patterns

In addition to the implicit prioritisation observed in their open-ended reasoning, experts were explicitly asked in Question 10 to rank the intentions of three stakeholders in the scenario. These stakeholders were the AV passenger, the cyclist, and the road policymakers. A rank of "1" indicated the highest priority. Table 2.5 summarises the aggregated rankings.

Table 2.5: Number of experts assigning each rank position to each stakeholder, based on their stated intentions in Question 10.

Stakeholder	**1st place**	**2nd place**	**3rd place**
Cyclist	12	5	0
Road policymakers	5	5	7
AV passenger	0	7	10

Out of 18 experts, one expert (ID02) declined to provide a ranking in Question 10, arguing that prioritising between stakeholders' intentions is inappropriate because legal and safety obligations should determine behaviour rather than preference-based trade-offs.

Thematic analysis of Question 11 justifications showed that the strong prioritisation of cyclists was grounded in their vulnerability as unprotected road users and in the moral obligation to minimise harm (for example, ID01, ID03, ID04, ID06, ID07, ID08, ID09, ID14, ID15, ID16, ID17, ID18). Several experts linked this priority to broader societal goals such as Vision Zero (ID09) and to the legal principle of protecting vulnerable road users (ID14). For these experts, safety considerations outweighed concerns for efficiency, trip time, or strict adherence to traffic regulations.

Experts who ranked road policymakers first (ID10, ID12, ID13) justified this choice by referring to the importance of systemic regulation and the rule of law in ensuring safe and predictable interactions for all road users. Some viewed policymakers as legitimate representatives of the public interest who are responsible for embedding safety into infrastructure and traffic rules (ID10, ID13). Others emphasised that the priorities of policymakers should align with the protection of vulnerable users, which indirectly supports cyclists (ID12).

The fact that no expert gave AV passengers the first rank was explained by their protected status within a vehicle and by the perception that their main concern, which is timely arrival, carries lower ethical weight compared to the safety of others (for example, ID06, ID08, ID09, ID11, ID16). When AV passengers were ranked second (for example, ID03, ID04, ID08, ID11, ID14, ID15, ID16), they were recognised as being directly involved in the situation. However, their needs were still seen as secondary to the safety of cyclists.

Overall, the explicit ranking task reinforces the patterns observed in the implicit analysis. Safety of vulnerable road users formed the primary decision criterion. Systemic safety considerations came second, and individual convenience came last.

2.4.5 Discussion

To address Objective 2, we applied the expert-derived reason categories developed in Section 3 to a context-specific case study involving a routine but ethically ambiguous AV scenario: overtaking a cyclist. This application contributes to **the theoretical gap** by operationalising the tracking condition of Meaningful Human Control (MHC) in everyday driving contexts, moving beyond high-stakes dilemmas to examine how AVs might align with human reasons in complex, real-world situations.

This section also speaks to **the methodological gap** by demonstrating how structured qualitative analysis—linking expert-elicited reasons to specific behavioural recommendations—can yield actionable insights. By capturing expert judgments on both expected (“will”) and preferred (“should”) AV behaviours, we identify how contextual factors, value tensions, and individual reasoning strategies shape nuanced expectations for AV decision-making. These insights form the basis for deriving a reason-based prioritisation principle and for developing an empirically grounded conceptual representation that maps how such reasons emerge and interact in context.

This focus on routine, ethically ambiguous scenarios expands on prior work that has largely centred on high-stakes dilemmas, such as crash scenarios or trolley problems (Rhim & Urban, 2021; Milford et al., 2025). Whereas those studies highlight binary moral choices, our study surfaces the nuance of everyday trade-offs and expert reasoning in context-rich decisions. To support the derivation and explanation of the reason-prioritisation principle, we developed a conceptual representation of reason-based AV decision-making for the cyclist-overtaking case (Figure 2.3) that illustrates how contextual factors give rise to reasons, how those reasons interact, and how prioritisation occurs in practice. This representation helps clarify the dynamics that underpin the prioritisation principle and demonstrates its applied relevance in ethically ambiguous everyday situations. While initially developed for the cyclist-overtaking scenario, its structure may also inform analyses of other routine ethically ambiguous driving contexts; however, further empirical investigation is required to examine such applicability.

Contextual Background Influencing Emerging Reasons. The reasons identified across expert responses consistently emerged from underlying contextual assumptions, including local regulations, technological capabilities, traffic situations, and individual differences. While the tracking condition in MHC theory requires that automated systems respond to human moral reasons (Santoni de Sio & Van den Hoven, 2018), existing literature does not explain how such reasons emerge from contextual circumstances. This study contributes a new insight by showing that expert reasons are shaped by situational background assumptions.

For instance, local regulations play a foundational role. One expert (ID13) stated that overtaking would be permissible under UK traffic laws. Conversely, experts ID7 and ID15 assumed stricter rules prohibiting overtaking, leading them to suggest manual takeover as a necessity. Technological capabilities and automation level also constrain or enable reasoning. Experts (IDs4, 15) noted that current AVs are programmed to avoid rule violations, thus requiring human intervention when overtaking is contextually necessary. This aligns with assumptions about Level 2 automation, where driver readiness remains essential. When experts assumed no manual override was available, they introduced alternative reasons such as *environmental impact*, *interaction management*, and *acceptance*.

Traffic situations, such as encountering a slow cyclist on a bidirectional road, influenced reasoning around safety, vigilance, and comfort. For instance, Expert ID5 emphasised that prolonged following increases driver workload, reinforcing the need for AVs to act pragmatically. Individual differences in expert values were also significant. Expert ID11 advocated strict rule-following as a matter of principle, while Expert ID 6 endorsed a consequentialist view, suggesting AVs should act based on outcomes (e.g., safety), regardless of legality.

Distinct Reasons Leading to Different Expected Behaviours. Distinct sets of reasons translate into differing AV behaviours. As shown in Figure 2.2, experts cited *rule compliance* and *safety* as dominant reasons for why AVs are expected to follow the cyclist. In contrast, overtaking behaviour was associated with a wider range of reasons, including *efficiency*, *comfort*, *acceptance*, and *interaction management*.

Three mechanisms were identified as key to understanding how reasons lead to behaviours. First, reason interpretation varies by context. *Safety*, for example, was interpreted as a reason to follow the cyclist (avoiding risky manoeuvres, ID4), but also to overtake (reducing traffic disruptions, ID17). This shows that a single reason can support opposing behaviours depending on situational framing.

Second, a tension between personal and collective interests was apparent. Personal motivations (e.g., arriving on time, IDs7, 15) often shaped expected behaviour. Meanwhile, collective values (e.g., environmental sustainability, ID17; shared road safety, ID 5) influenced preferred actions. This tension reflects a broader ethical divide in AV decision-making, as noted in prior work (Bonnefon et al., 2016).

Third, balancing multiple reasons emerged frequently. Experts combined several motivations in a single rationale (e.g., ID 6 cited both safety and comfort). Notably, when rule compliance conflicted with more practical or safety-oriented reasons, the latter were often prioritised. These dynamic prioritisation mechanisms echo the tracking model proposed by Mecacci & Santoni de Sio (2020), in which AV systems are designed to respond to the most proximal reasons unless overridden by more distal values. Our findings complement this by empirically showing

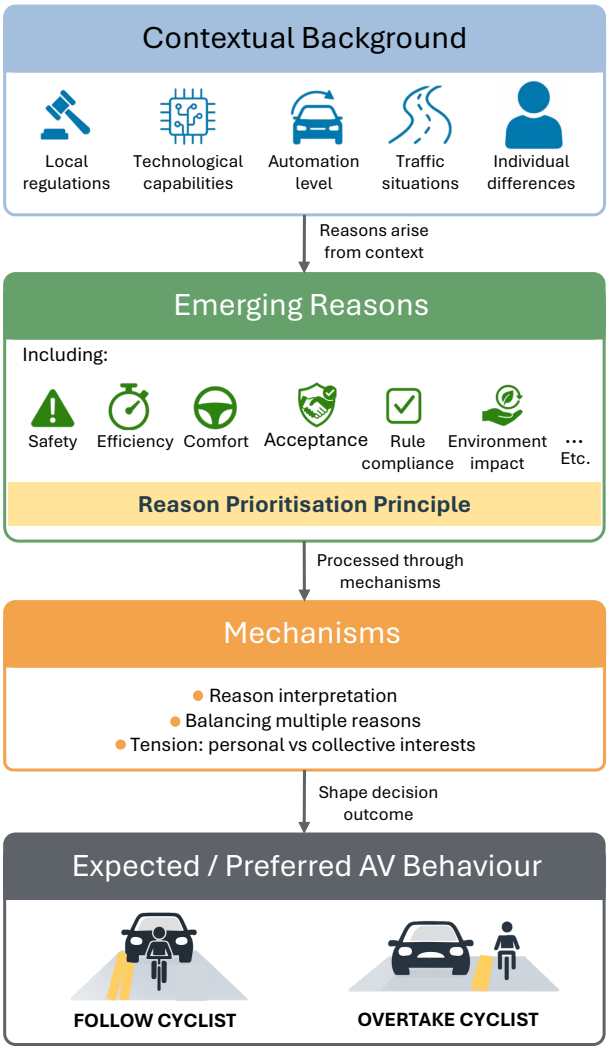


Figure 2.3: Conceptual representation of reason-based AV decision-making in the cyclist-overtaking scenario. The representation illustrates how contextual background factors give rise to reasons for AV behaviour and where the derived reason-prioritisation principle operates within the AV decision-making process. These reasons, together with their prioritisation, are processed through mechanisms such as reason interpretation, tensions between personal and collective interests, and the balancing of multiple reasons. Collectively, these dynamics shape expected and preferred AV behaviour when deciding whether to follow or overtake a cyclist in an ethically ambiguous everyday situation.

that human experts interpret and prioritise reasons fluidly, often allowing situational context to shape whether a reason like safety or legality dominates. This underscores the challenge of implementing fixed hierarchies of reasons in AV design and supports the need for flexible, context-sensitive reason-tracking mechanisms.

Implicit and Explicit Prioritisation Among Reasons. While experts cited distinct reasons for AV behaviour, a clear prioritisation pattern emerged across both implicit and explicit analyses. In their open-ended reasoning, *safety* consistently served as the primary, non-negotiable consideration. Other reasons, such as *efficiency*, *comfort*, *interaction management*, and *acceptance* appeared were frequently mentioned but typically framed in relation to safety, making them secondary. *Rule compliance* was treated as conditional obligation: important when it aligned with safety and traffic flow, but often overridden when it did not, particularly when deviations could serve other high-priority values without compromising safety.

The explicit ranking reinforced this structure. Cyclists, as the most vulnerable road users in the scenario, were ranked first by majority of experts and never ranked last. AV passengers were never ranked first, most often placed last, and seen as holding lower priority due to their protected status inside the vehicle. Road policymakers occupied a middle position, valued for their role in setting systemic safety rules, but secondary to the immediate protection of vulnerable road users.

Together, these findings indicate that experts expect AVs to prioritise the safety of the most vulnerable above all else, followed by broader public and systemic safety, and lastly the convenience of more protected individuals such as AV passengers. Secondary values like comfort, efficiency, and environmental impact were seen as important, but acceptable only when they did not conflict with the safety of vulnerable users.

This prioritisation structure aligns with previous findings in AV ethics literature, where safety is widely regarded as the paramount consideration. For instance, the Moral Machine experiment (Awad et al., 2018) and expert-based studies (Milford et al., 2025) show broad consensus that risk minimisation should guide AV behaviour. Similarly, rule compliance has been treated as a conditional obligation in earlier work, particularly when rigid adherence may compromise safety or efficiency (Ma & Feng, 2024).

However, our results extend this discussion by demonstrating that such prioritisation patterns are not limited to high-stakes dilemmas but also emerge in routine, ethically ambiguous situations. This suggests that value trade-offs are not confined to emergencies, but are an ongoing feature of everyday AV operations. Moreover, while existing frameworks often rely on normative claims or abstract models of ethical reasoning, our findings reveal how experts actually balance and contextualise competing considerations—including legality, comfort, and environmental impact—based on real-world constraints and assumptions. This contributes a more fine-grained, empirically grounded understanding of how prioritisation unfolds across different layers of AV behaviour, and how certain reasons rise or recede in relevance depending on the driving context.

Principle of Reason Prioritisation in the Overtaking Cyclist Scenario. Building on these prioritisation patterns, we derive a principle of reason prioritisation in ethically ambiguous routine situations, particularly in the overtaking cyclist scenario. The synthesis of expert reasoning suggests three core guidelines.

First, AVs must always prioritise safety of the most vulnerable road users, such as cyclists in this scenario, over the interests of more protected users, including AV passengers.

Second, legal compliance should be the default behaviour. However, when strict rule-following would conflict with the safety of vulnerable users, or when it would go against collective interests such as avoiding discomfort, confusion, or indirect risks to road users and broader public, then carefully constrained deviations may be justified-provided that safety can be fully maintained.

Third, any justified deviation from traffic rules should be aimed at serving the greater public good rather than the convenience of the AV's occupants alone.

Unlike previous models that focus on rare, high-stakes scenarios—such as the MIT Moral Machine's global crash dilemma study (Awad et al., 2018), or algorithmic approaches like Augmented Utilitarianism (AU) (Gros et al., 2025b)—our principle addresses a less examined but highly relevant domain: ethically ambiguous, everyday driving situations.

The Moral Machine revealed diverse cultural preferences in life-or-death crash dilemmas, underscoring the challenge of creating globally acceptable AV ethics. However, its emphasis on binary, extreme scenarios limits its applicability to nuanced real-world contexts. Similarly, AU advances ethical reasoning by incorporating diverse moral theories into adaptable goal functions grounded in empirical data. It uses attributes like harm, fairness, and legality, refined through participatory methods, to compute ethical decisions dynamically. In their work, Gros et al. (2025b) designed AU to address both critical and non-critical contexts. However, the scenario used to illustrate and evaluate AU, such as brake-failure dilemmas or unusual narrow-road positioning with vulnerable pedestrians, are still relatively rare compared to the more commonplace, low-stakes situations AVs regularly encounter in everyday driving situations such as overtaking a slow cyclist.

In contrast, our principle complements these by providing a practice-orientated structure for routine AV behaviour, grounded in expert judgement rather than abstract moral theory or crowdsourced moral preferences. It emphasises prioritising the safety of the most vulnerable, allowing pragmatic flexibility when safety is not a risk, and applying a public-good focus when making legal exceptions. This structure captures the complex value trade-offs that arise in cyclist overtaking scenarios in everyday AV operation. While these decisions may not involve stark life-or-death choices, they can meaningfully contribute to the development of AVs that are ethically designed to build public trust and acceptance.

Further, in addition to its relevance to everyday rather than exceptional moral dilemmas, our prioritisation principle has potential relevance to computational implementation. Prior work has shown that human-provided reasons can be operationalised by quantifying them through human-factors research and embedding them, with associated weights, into a trajectory evaluation framework to handle ethically challenging routine driving scenarios (Suryana et al., 2025c). This quantification and weighting structure provide a possible pathway for integrating our reason categories and prioritisation principle into future evaluation stages of trajectory planning.

Recommendations and Implications. The recommendations presented here apply directly to overtaking situations and may be applicable to similar ethically ambiguous driving scenarios—contexts in which AVs must navigate tensions between rule adherence, safety, comfort,

and public expectations. Our findings on contextual reason prioritisation framework suggest that AV systems should incorporate flexible, context-aware decision logic. Developers should embed mechanisms to interpret reasons dynamically and prioritise them based on real-time conditions. Policymakers should consider enabling AVs to operate within regulated bounds that allow principled flexibility—particularly in routine scenarios where rigid rule-following may be counterproductive. Designers should also consider how human-like behaviour and acceptance intersect with safety and efficiency. Expert ID 11 highlighted that users are more likely to trust AVs that behave like human drivers, provided that safety is preserved. This has implications for user-centred AV design and for the development of regulatory frameworks that accommodate safe and socially acceptable deviations.

Limitations and Future Directions. This framework, while grounded in rich expert input, has limitations. Expert reasoning may reflect regional or cultural biases, and its generalisability across different AV scenarios—such as urban versus rural environments, or contexts with varying cultural norms—remains untested.

A further limitation concerns the interpretative nature of the implicit prioritisation analysis. Although the study included an explicit prioritisation task in which experts directly ranked the importance of reasons, the implicit prioritisation structure was inferred through qualitative interpretation of expert explanations. While independent coding and consensus resolution helped mitigate potential bias, these processes reduce rather than eliminate subjectivity. Future research could incorporate larger-scale empirical validation to examine how the prioritisation structure generalises beyond expert accounts.

Additionally, future research could examine how human reasons evolve over time in real or simulated driving contexts, and how AVs might adapt their decision-making while maintaining the safety of vulnerable road users and enabling practical action in situations where strict rule compliance conflicts with other considerations. While prior work (Suryana et al., 2025c) demonstrates the feasibility of operationalising human-provided reasons by quantifying them within a trajectory-evaluation framework, this work addresses the evaluation of candidate trajectories rather than full real-time control. The integration of the reason-prioritisation structure proposed in the present study into such computational frameworks has not yet been implemented or validated. Future work should develop and test this integration in both simulation and controlled real-world scenarios to assess practical effectiveness and feasibility.

2.5 Conclusion

This study derives a reason-prioritisation principle from expert reasoning about cyclist overtaking in ethically ambiguous routine driving situations. Grounded in the tracking condition of Meaningful Human Control (MHC), the principle supports an expert-derived framework that maps how human reasons influence AV decisions. Through qualitative interviews with AV experts, we identified thirteen categories of reasons that influence manoeuvre planning, structured across normative, strategic, tactical, and operational levels of AV behaviour, and linked to the roles of relevant human agents.

The findings show that AV decisions often involve multiple overlapping reasons, with *safety* consistently regarded as the primary concern. Other reasons, such as *efficiency*, *comfort*, and

acceptance, were frequently mentioned alongside safety but rarely overrode it. *Rule compliance* was treated as a conditional obligation and often deprioritised when it conflicted with more context-sensitive goals. These prioritisation patterns a set of empirically grounded principles that upholds safety while permitting carefully constrained deviations from legal rules when justified by practical or ethical considerations.

By mapping expert reasons into a layered structure and positioning them within a conceptual representation of reason-based AV decision-making, this case-specific model offers guidance that complements existing high-stakes ethical approaches and may inform future research on AV behaviour in dynamic, real-world scenarios. Future work should evaluate the broader applicability of this representation across diverse cultural contexts, automation levels, and driving situations.

Chapter 3

Formal Representation of Human Reasons and Supervisory Framework for Behavioural Adjustment

---

This chapter develops a formal representation of human reasons within automated vehicle (AV) decision-making frameworks and implements a supervision framework that builds on this representation. Building on the empirical findings of Chapter 2, it translates categories such as safety, efficiency, and regulatory compliance into quantitative models that can be operationalised in real time. These models form the basis of a human reasons-based supervision framework that monitors the alignment between AV behaviour and human reasons and triggers behavioural adjustments when misalignment is detected. The chapter further examines the framework's behaviour in simulated overtaking scenarios and evaluates its ability to enable a balance among safety, efficiency, and regulatory compliance in ethically ambiguous traffic situations.

This chapter presents a single contribution derived from the following paper, published at the 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS):

Suryana, L. E., Rahmani, S., Calvert, S. C., Zgonnikov, A., and van Arem, B. (2025). A Human Reasons-Based Supervision Framework for Ethical Decision-Making in Automated Vehicles. In Proceedings of IROS 2025 (pp. 21495–21502). IEEE.

3.1 Abstract

Ethical dilemmas are a common challenge in everyday driving, requiring human drivers to balance competing priorities such as safety, efficiency, and rule compliance. However, much of the existing research in automated vehicles (AVs) has focused on high-stakes ”trolley problems,” which involve extreme and rare situations. Such scenarios, though rich in ethical implications, are rarely applicable in real-world AV decision-making. In practice, when AVs confront everyday ethical dilemmas, they often appear to prioritise strict adherence to traffic rules. By contrast, human drivers may bend the rules in context-specific situations, using judgement informed by practical concerns such as safety and efficiency. According to the concept of meaningful human control, AVs should respond to human reasons, including those of drivers, vulnerable road users, and policymakers. This work introduces a novel human reasons-based supervision framework that detects when AV behaviour misaligns with expected human reasons to trigger trajectory reconsideration. The framework integrates with motion planning and control systems to support real-time adaptation, enabling decisions that better reflect safety, efficiency, and regulatory considerations. Simulation results demonstrate that this approach could help AVs respond more effectively to ethical challenges in dynamic driving environments by prompting replanning when the current trajectory fails to align with human reasons. These findings suggest that our approach offers a path toward more adaptable, human-centered decision-making in AVs.

3.2 Introduction

Addressing the ethical complexities that emerge in daily driving contexts remains essential for social acceptance of automated vehicles (AVs). Despite their promised advantages in safety improvements and transportation access (Geisslinger et al., 2021), the widespread adoption of these systems hinges on their capacity to reflect human ethical judgment, particularly when confronting morally ambiguous situations where multiple values compete (Lin, 2016; Millar et al., 2017)—situations commonly referred to as ethical dilemmas. Examples include deciding whether to briefly occupy the opposite lane to safely overtake cyclists (FSDEvolution, 2025), or speeding up temporarily to avoid unsafe situations. This behaviour reveals a critical gap in AV decision-making: the necessity of designing AVs capable of dynamically balancing multiple considerations, such as safety, efficiency, regulatory compliance, and contextual appropriateness, in real-time, rather than relying solely on predefined regulations.

Current approaches show limitations when it comes to handling everyday ethical dilemmas in automated driving. Much of the research has focused on extreme scenarios, such as the well-known “trolley problem” (Bonnefon et al., 2019). While philosophically significant, trolley problems rarely happen in daily driving, and despite the practical importance of everyday dilemmas, they often receive less attention (Geisslinger et al., 2021; Himmelreich, 2018). (Lin, 2016) emphasises that everyday ethical decisions in automated driving extend far beyond these extreme scenarios, requiring nuanced contextual use of reasons, something current systems lack. Similarly, (Nyholm & Smids, 2016) argue that framing AV ethics as simplified trolley problems fails to capture the probabilistic nature and dynamic complexity of real-world driving situations.

Although the ethical dimensions of “mundane” scenarios may seem straightforward for hu-

man drivers, they require context-aware judgment that comes naturally to human but poses challange for AVs. These judgments must balance multiple ethically relevant considerations, such as safety, efficiency, and social norms. This adaptability represents a challenge for AV systems designed with traditional motion planning algorithms, which primarily optimise for trajectory smoothness and collision avoidance without explicitly integrating ethical considerations. Building on this understanding, recent ethical frameworks propose more holistic approaches that better align with human moral intuitions. (Cecchini et al., 2024) and (Henschke, 2020) collectively emphasise that effective AV ethics must integrate moral principles such as deontological ethics, virtue ethics, and consequentialist considerations while maintaining transparency in decision-making.

However, integrating these ethical principles into AV decision-making remains challenging. While prior works have focused on embedding such principles into control and motion algorithms (Geisslinger et al., 2021; Thornton et al., 2016; Geisslinger et al., 2023), current approaches fail to make explicit when these principles are in conflict. Recognising these conflicts is essential for enabling transparent decisions and for adjusting AV behaviour to better reflect the ethical principles the system is intended to uphold.

The concept of meaningful human control (MHC) (Mecacci & Santoni de Sio, 2020; Santoni de Sio & Van den Hoven, 2018) offers a promising conceptual bridge for addressing the challenge of making explicit which moral principles are in conflict. MHC asserts that humans should ultimately be responsible for every decision made by automated systems. (Santoni de Sio & Van den Hoven, 2018) laid the groundwork for achieving MHC. One of the required conditions is the tracking condition, which requires automated systems to respond to the reasons of relevant humans. In the remainder of this paper, we refer to these relevant humans – such as drivers, passengers, pedestrians, and policymakers – as stakeholders. According to Mecacci & Santoni de Sio (2020), these reasons can be understood as moral values or principles that are reflected in human driving plans and intentions—such as ensuring safety and comfort for both themselves and others, driving efficiently, and complying with traffic regulations. From this perspective, if an AV is designed to uphold certain moral principles, the tracking condition provides a clear expectation that its behaviour should reflect corresponding human plans and intentions.

To operationalise this concept and address the challenges of handling ethical dilemmas and making moral principles explicit in AV decision-making, we propose a novel human reasons-based supervision framework that enables AVs to evaluate if their behaviour aligns with the reasons of diverse stakeholders. By grounding this framework in the tracking condition of meaningful human control, we aim to support AV decision-making in ethically challenging everyday scenarios that require balancing multiple, sometimes conflicting, values.

Specifically, the primary contribution of this paper is a modular human reasons-based supervision framework that enables AVs to make ethically nuanced decisions in routine yet ethically challenging scenarios. The framework continuously evaluates how well the AV's behaviour aligns with human reasons and triggers replanning when a misalignment is detected. The paper contributes:

1. We developed a detection mechanism that uses stakeholder reason scores and predefined thresholds to identify when AV behaviour misaligns with human reasons;

2. We integrated the human reasons-based supervision framework into an AV control archi-

tecture, including a mechanism for triggering replanning when reason scores fall below predefined thresholds;

3. We enabled explainability by using reason scores as interpretable indicators of why behaviour changes are recommended in routine, ethically challenging situations.

This work advances the discourse on AV decision-making in ethically challenging transportation scenarios by bridging the gap between moral principles and practical AV decision-making, ultimately supporting the development of socially acceptable automated mobility solutions that align with human reasons across diverse everyday scenarios.

The remainder of this paper is organised as follows: Section 3.3 presents the detailed methodology and system architecture, including the mathematical human reasons and its integration into a motion planning framework. Section 3.4 describes the experimental setup and simulation environment. Section 3.5 and 3.6 present and discuss the simulation results and the impact of ethical supervision on vehicle behaviour. Finally, Section 3.7 concludes the paper and outlines directions for future research.

3.3 Methodology

3.3.1 Problem Formulation

We formalise the automated vehicle navigation problem in scenarios involving ethical decision-making. While the framework is generic and applicable to a wide range of situations, for demonstration purposes, we consider a scenario including the interaction of an automated vehicle with a vulnerable road user (VRU). In this scenario, an automated vehicle navigating a bidirectional road faces an ethical dilemma during overtaking manoeuvres. To maintain efficient travel, the vehicle must either follow the VRU with a very low speed, which is not desirable for the vehicle's passenger, or overtake a slower-moving VRU, which may require temporarily entering the oncoming lane or reducing the safety buffer with the cyclist. This manoeuvre challenges forces a trade-off between strict compliance, user safety, and travel efficiency.

For the problem formulation, we consider the automated vehicle operating in state space $\mathcal{X} \subset \mathbb{R}^n$ with state vector $\mathbf{x_t} = [\mathbf{p_t}, v_t, \theta_t, \omega_t]^T$, where $\mathbf{p_t} = [p_x, p_y]^T$ represents position, v_t denotes velocity, θ_t is heading angle, and ω_t is rotational velocity. The control space $\mathcal{U} \subset \mathbb{R}^m$ consists of $\mathbf{u_t} = [a_t, \delta_t]^T$, representing acceleration and steering angle.

Our multi-agent ethical framework defines stakeholders $\mathcal{S} = \{s_1, s_2, ..., s_k\}$ with reason functions $R_{s_i} : \mathcal{X} \times \mathcal{U} \to [0,1]$ quantifying alignment with each stakeholder's perspective. This formulation means that each function R_{s_i} evaluates the vehicle's state and control actions to produce a score between 0 and 1, reflecting how well the AV's behaviour satisfies the ethical priorities of the respective stakeholder. For the designed scenario, we identify three key stakeholders: road policymakers (s_{policy}), vulnerable road user (s_{VRU}), and drivers (s_{driver}).

The navigation problem for the automated vehicle is formulated as:

$$\begin{aligned} \min_{u_0,\ldots,u_{T-1}} \quad & \sum_{t=0}^{T-1} \mathcal{J}(\mathbf{x_t}, \mathbf{u_t}) \\ \text{s.t.} \quad & \mathbf{x_{t+1}} = f(\mathbf{x_t}, \mathbf{u_t}) \\ & \mathbf{x_t} \in \mathcal{X}_{safe} \\ & \mathbf{u_t} \in \mathcal{U} \\ & R_{s_i}(\mathbf{x_t}, \mathbf{u_t}) \geq \tau_{s_i}, \forall s_i \in \mathcal{S} \end{aligned} \tag{3.1}$$

Here, $\mathcal{J}(\mathbf{x_t}, \mathbf{u_t})$ represents the cost function to be minimised over the control horizon T, evaluating the performance of the AV's state and control inputs at each time step t. The function $f(\mathbf{x_t}, \mathbf{u_t})$ denotes the system dynamics model that predicts the next state $\mathbf{x_{t+1}}$ based on the current state $\mathbf{x_t}$ and control input $\mathbf{u_t}$. The set $\mathcal{X}_{safe} \subset \mathcal{X}$ defines the safe region of the state space where the vehicle must operate to avoid collisions and other hazards. Finally, for each stakeholder s_i, τs_i denotes the threshold value specifying the minimum acceptable reason score that the AV's behaviour must meet.

3.3.2 Framework Architecture

Our approach implements a multi-component and hierarchical framework with three main components:

1. *Global Motion Planning*: Responsible for finding a reference trajectory for the vehicle to be followed. It uses A* search with motion primitives to generate feasible reference paths from the current state of the vehicle to the goal location.

2. *Model Predictive Control*: Optimises vehicle trajectory when following the reference path. It ensures kinodynamic feasibility and satisfying soft and hard constrained defined, such as safety, efficiency, and comfort.

3. *Human Reasons-based Supervision Framework*: Evaluates the planned actions against ethical criteria and triggers replanning when necessary if the criteria are not met.

These elements are depicted in Fig. 3.1. The key innovation in our approach is the definition and integration of a human reasons-based supervision framework as a mechanism for triggering replanning, ensuring that the vehicle's behaviour satisfies ethical constraints derived from multiple stakeholders' perspectives. Therefore, we begin by detailing the components of the framework.

3.3.3 Human Reasons-based Supervision Framework

Our approach to developing a framework that supervises the alignment between AV behaviour and human reasons builds on the qualitative evaluation steps for the tracking condition outlined by (Suryana et al., 2025b). These steps involve defining the relevant stakeholders and articulating their reasons, as well as specifying the features of the AV system that govern its behaviour.

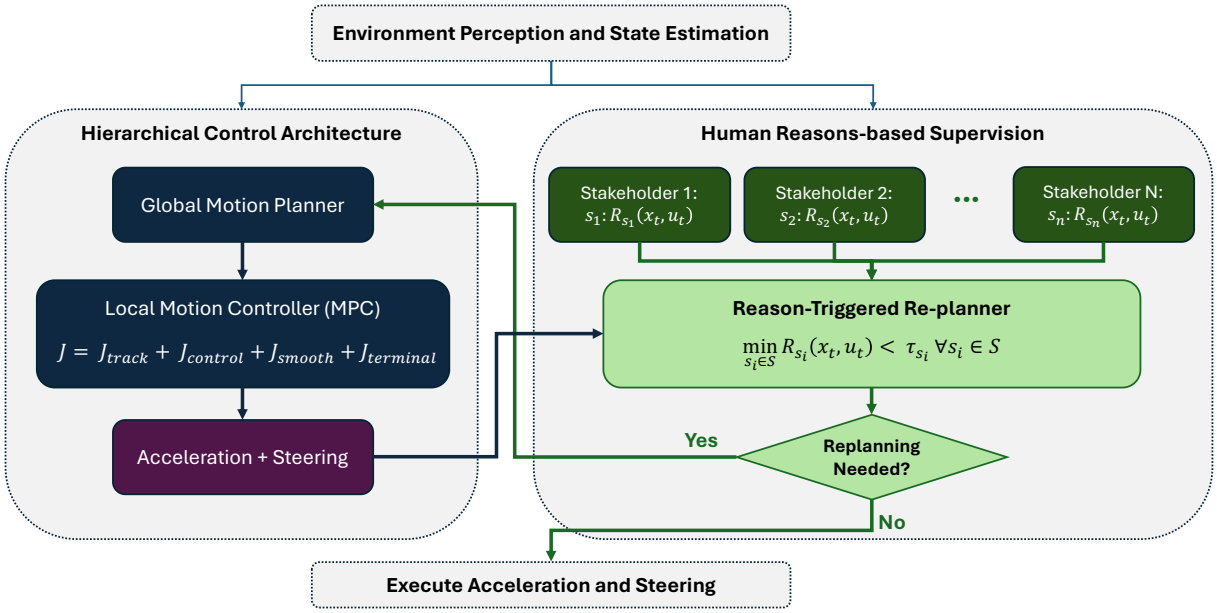


Figure 3.1: Hierarchical control architecture with human reasons-based supervision for ethical AV decision-making

While the original approach remains qualitative, our work extends it by developing a quantitative framework. Specifically, after identifying the stakeholders, we formalise their reasons mathematically. This process is described in this section, while the features of the AV system that govern its behaviour are presented in Section 3.3.4.

Identification of Stakeholders and Reason Models

To effectively integrate human ethical considerations into AV decision-making, it is essential to identify the key stakeholders involved and define how system's behaviour influence their alignment with stakeholders' reasons. Accordingly, we model three primary stakeholders in the designed scenario:

- *Road Policymaker*: Represents regulatory authorities whose reason is to ensure overall road safety through regulatory compliance.
- *Vulnerable Road User*: Represents vulnerable road users whose reason is to commute with safety and comfort.
- *Driver*: Represents the vehicle occupant's motivation for efficiency and arriving at the destination as fast as possible.

To operationalise stakeholder perspectives within our framework, we establish specific reason models for each. For each stakeholder, we define a reason model that quantifies their satisfaction with the vehicle's behaviour on a scale from 0 to 1, where 1 indicates full satisfaction and 0 indicates complete dissatisfaction.

Mathematical Representation of Reason Models

A key contribution of our framework is the operationalisation of abstract human reasons using empirically supported, measurable parameters. While previous work has discussed human reasons conceptually (Suryana et al., 2025b) or proposed initial variables (Calvert & Mecacci, 2020), these efforts have not established an empirically grounded mapping to real-world driving variables. We address this gap by introducing a reason model grounded in human factors research.

To translate stakeholder's reason into a computationally viable form, we adopt a set of piecewise exponential functions that model how stakeholder satisfaction declines when specific behavioural threshold are crossed. These modelling choice are grounded based on both computational simplicity and their ability to approximate human reasons. As demonstrated by Tversky & Kahneman (1992), individuals tend to overweight low probabilities and underweight high probabilities, implying a rapid shift in perceived risk once certain thresholds are crossed. Thus, the exponential function is chosen because it can well capture the rapid change in perceived acceptability. Nevertheless, the proposed equations serve as a representative model that can be adjusted via thresholds and scaling constants to suit various scenarios.

To ensure realism, each stakeholder's reason is modelled using scenario-specific variables informed by human factors research. Details of the experimental case appear in Section 3.4. For example, the cyclist's reason—related to comfort and perceived safety—is represented using lateral distance and tailgating time, based on findings from Road Safety Authority (2018); Oskina et al. (2023). Here, comfort refers to the cyclist's subjective experience of emotional and physical ease during interaction with the vehicle. Empirical studies show that insufficient lateral clearance and prolonged close following increase stress and perceived risk, justifying the use of these variables as proxies.

The same vehicle behaviour–prolonged following–also impact driver's reason, which relates to driving efficiency. This is modelled as perceived impatience, based on time and distance the AV follows the cyclist at close range. As shown in Lee (2010), such conditions could lead to frustration under time pressure. Meanwhile, the policymaker's reason—regulatory compliance—is satisfied when the AV stays within its designated lane. The reason's score decreases as the AV crosses into the opposite lane to overtake the cyclist. This reflects findings from Suryana et al. (2025a), where experts preferred the AV to overtake gently and return to its lane promptly, indicating that in such situations, any lane violation should be considered minimal.

Policymaker's Reason The policymaker's reason score quantifies regulatory compliance – specifically, adherence to lane regulations in this scenario:

$$R_{\text{policymaker}} = \begin{cases} 1, & \text{if } d_{\text{veh}} > 0, \\ e^{k_1 \cdot d_{\text{veh}}}, & \text{otherwise.} \end{cases} \tag{3.2}$$

where d_{veh} is the lateral displacement of the ego vehicle from the center line (positive when on the correct side of the road, negative when in the oncoming lane); k_1 is a scaling constant.

VRU's Reason The VRU's reasons score is decomposed into safety assurance and comfort preservation.

- *Safety Assurance:*

$$R_{sa}(t) = \begin{cases} 1, & \text{if } d_{\text{veh-vru}} > d_{\text{th,vru}}, \\ \frac{1}{e^{k_2 (d_{\text{veh-vru}} - d_{\text{th,vru}})}}, & \text{otherwise}, \end{cases} \tag{3.3}$$

where $d_{\text{veh-vru}}$ denotes the distance between the vehicle and the VRU; $d_{\text{th,vru}}$ is the perceived safe distance threshold; k_2 is a scaling constant.

- *Comfort Preservation:*

$$R_{cp}(t) = \begin{cases} 1, & \text{if } t_{\text{close,vru}} < t_{\text{th,vru}} \text{ or} \\ & d_{\text{veh-vru}} > d_{\text{th,vru}}, \\ \frac{1}{e^{k_3 (t_{\text{follow}} - t_{\text{th,vru}})}}, & \text{otherwise}. \end{cases} \tag{3.4}$$

where $t_{\text{close,vru}}$ is the cumulative time during which $d_{\text{veh-vru}} < d_{\text{th,vru}}$; $t_{\text{th,vru}}$ is the maximum tolerable time for the cyclist to be followed too closely; k_3 is a scaling constant.

The overall VRU's reason score is then given by:

$$R_{\text{VRU}}(t) = R_{sa}(t) \cdot R_{cp}(t). \tag{3.5}$$

where $R_{\text{VRU}}(t)$ combines the safety and comfort components.

Driver's Reason The driver's reason score is defined as:

$$R_{\text{driver}}(t) = \begin{cases} 1, & \text{if } t_{\text{behind,driver}} < t_{\text{th,driver}} \text{ or} \\ & d_{\text{veh-vru}} > d_{\text{th,driver}}, \\ \frac{1}{e^{k_4 (t_{\text{behind,driver}} - t_{\text{th,driver}})}}, & \text{otherwise}, \end{cases} \tag{3.6}$$

where $t_{\text{behind,driver}}$ is the cumulative time during which $d_{\text{veh-vru}} < d_{\text{th,driver}}$; $t_{\text{th,driver}}$ is the time threshold for close following that the driver considers acceptable; $d_{\text{veh-vru}}$ is the distance between the vehicle and the VRU; $d_{\text{th,driver}}$ is the distance threshold below which the driver considers the AV to be following too closely, leading to perceived inefficiency; and k_4 is a scaling constant. Note that the scores of k_1, k_2, k_3, and $k_4 = 0.2$ in our experiment can be adjusted depending on how quickly we want the reasons to shift from 1 to 0 when the reason thresholds are crossed.

3.3.4 Motion Planning and Control Implementation

To demonstrate the generalisability and practical implementability of our framework, we integrate the human reason-based supervision framework into an AV feature that governs its behaviour. In this research, we adopt an established motion planning and control framework (Rahmani et al., 2023). In the following, we briefly describe the underlying motion planning and control mechanism and explain how the reason-triggered replanning seamlessly fits into this structure.

Motion Planning

The motion planner in this study builds a directed graph from the vehicle's current state using pre-computed motion primitives. An A* algorithm is applied to find the optimal path by minimizing the cost function

$$\begin{aligned} J_{\text{path}} =& w_1 \cdot J_{\text{length}} + w_2 \cdot J_{\text{smoothness}} \\ &+ w_3 \cdot J_{\text{obstacle\_clearance}} + w_4 \cdot J_{\text{traffic\_rule}} \end{aligned} \tag{3.7}$$

where J_{length} is the cost related to the length of the path from the initial state to the goal state, aiming to motivate the shortest path; $J_{\text{smoothness}}$ is the cost related to the smoothness of the path; $J_{\text{obstacle\_clearance}}$ is the cost for avoiding obstacles; and $J_{\text{traffic\_rule}}$ is the cost aiming to avoid areas prohibited by traffic rules. The output of the planner is a reference trajectory passed to the controller for execution. We employ a modified version of A* to enhance search efficiency and applicability for our specific use case. The detailed algorithm implementation is documented in (Rahmani et al., 2025). It's worth noting that the cost function has been slightly modified to fit the purpose of this study. More specifically, the weights related to the costs for obstacle clearance and traffic adherence have been separated compared to the standard implementation in (Rahmani et al., 2025).

Controller

We formulate a finite-horizon optimisation problem that is solved at each time step to ensure trajectory following while respecting user-defined constraints.

The vehicle state at the time step t is represented as:

$$\mathbf{x}(t) = [x_t, y_t, \theta_t, v_t]^T \tag{3.8}$$

comprising position coordinates (x_t, y_t), heading angle θ_t, and longitudinal velocity v_t. The control inputs are:

$$\mathbf{u}(t) = [a_t, \delta_t]^T \tag{3.9}$$

where a_t denotes acceleration and δ_t the steering angle.

Vehicle dynamics are modelled using a bicycle model as follows:

$$\dot{x} = v\cos(\theta+\beta), \dot{y} = v\sin(\theta+\beta), \dot{\theta} = \frac{v}{L}\sin(\beta), \dot{v} = a \tag{3.10}$$

with slip angle $\beta = \arctan(\frac{l_r}{L}\tan(\delta))$, wheelbase L, and rear axle distance l_r. We discretise this continuous model using time step T_s:

$$\mathbf{x}(t+1) = A_d\,\mathbf{x}(t) + B_d\,\mathbf{u}(t) + \mathbf{d}_d \tag{3.11}$$

where the discrete system matrices are:

$$A_d = \begin{bmatrix} 1 & 0 & T_s c_\theta & -T_s v_t s_\theta \\ 0 & 1 & T_s s_\theta & T_s v_t c_\theta \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \dfrac{T_s \tan(\delta_t)}{L} & 1 \end{bmatrix} \tag{3.12}$$

with $c_\theta = \cos(\theta_t)$ and $s_\theta = \sin(\theta_t)$ for brevity. The input matrix and disturbance term are:

$$B_d = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ T_s & 0 \\ 0 & \dfrac{T_s v_t}{L\cos^2(\delta_t)} \end{bmatrix} \tag{3.13}$$

$$\mathbf{d}_d = \begin{bmatrix} T_s v_t s_\theta \theta_t \\ -T_s v_t c_\theta \theta_t \\ 0 \\ -\dfrac{T_s v_t \delta_t}{L\cos^2(\delta_t)} \end{bmatrix} \tag{3.14}$$

Our cost function integrates multiple objectives over prediction horizon N:

$$\begin{aligned} J = &\sum_{t=0}^{N-1} (\|e_t^\perp\|_{Q_\perp}^2 + \|e_t^\parallel\|_{Q_\parallel}^2 + \|e_{\theta v,t}\|_{Q_{\theta v}}^2) \\ &+ \sum_{t=0}^{N-1} (\|\mathbf{u}_t\|_R^2 + \|\Delta\mathbf{u}_t\|_{R_d}^2) \\ &+ \|\mathbf{x}_N - \mathbf{x}_{ref,N}\|_{Q_f}^2 \end{aligned} \tag{3.15}$$

where $e_t^\perp$ and $e_t^\parallel$ represent perpendicular and parallel trajectory tracking errors, $e_{\theta v,t}$ captures orientation and velocity errors, and $\Delta\mathbf{u}_t = \mathbf{u}_{t+1} - \mathbf{u}_t$. The matrices $Q_\perp$, $Q_\parallel$, $Q_{\theta v}$, R, R_d, and Q_f are weighting matrices that prioritize different aspects of performance.

The optimisation operates under constraints:

$$\mathbf{x}_{t+1} = f(\mathbf{x}_t, \mathbf{u}_t); \mathbf{u}_t \in \mathcal{U}; \mathbf{x}_t \in \mathcal{X} \tag{3.16}$$

where $\mathcal{U}$ defines input limitations:

$$a_{min} \leq a_t \leq a_{max}; \delta_{min} \leq \delta_t \leq \delta_{max} \tag{3.17}$$

Reason-Triggered Replanning

At each time step, the system evaluates the reason scores for all stakeholders using the formulations provided in Eq. 3.2 to Eq. 3.6. For each stakeholder $s_i \in \mathcal{S}$, the reason score $R_{s_i}(\mathbf{x_t}, \mathbf{u_t})$ is computed. If any score falls below its corresponding threshold τ_{s_i},

$$\min_{s_i \in \mathcal{S}} R_{s_i}(\mathbf{x_t}, \mathbf{u_t}) < \tau_{s_i}, \tag{3.18}$$

the system immediately triggers a replanning cycle. During replanning, the current scenario is updated, and a new reference trajectory is generated and passed to the controller. This continuous evaluation ensures that the vehicle's motion remains aligned with the ethical and performance criteria of all stakeholders. In our implementation, the path finding algorithm incorporates a set of weights in its cost function (Eq. 3.7) to provide flexibility when replanning is needed. For instance, under normal circumstances, prohibited areas by traffic rules are treated similarly to obstacles by assigning large weights to the $J_{\text{traffic\_rule}}$, restraining the A* search algorithm from generating paths through those areas. When replanning is triggered due to misalignment with human reasons, prohibited areas could temporarily receive lower costs, allowing the A* algorithm to search through those areas and provide a new trajectory with a different, potentially higher, human-reasoning score. Since the replanning strategy is not within the scope of this study, we refer the readers to the implementation of our planner detailed in (Rahmani et al., 2025). We would like to highlight that the proposed evaluation framework remains algorithm-agnostic and can assess trajectories generated by any motion planning approach.

3.4 Experiment Setup

To evaluate our human reasons-based supervision framework, we test it in an ethically challenging cyclist overtaking scenario, where an ego vehicle traveling in the right lane encounters a slow-moving cyclist on a narrow road (Fig. 3.2). Safely overtaking requires the vehicle to briefly enter the left lane, which is normally reserved for oncoming traffic. This forces a trade-off between strict lane adherence and efficient, safe manoeuvring, highlighting the ethical dilemma arising from the conflicting priorities of the involved stakeholders:

- *Road Policymakers* enforce traffic regulations that prohibit left-lane usage to ensure overall road safety.
- *Cyclists* require a safe and comfortable riding experience, which may be compromised by vehicles manoeuvring too closely.
- *Drivers* aim for efficient travel, potentially pressuring the system to overtake despite the inherent safety and regulatory concerns.

The detailed definitions of the reason models for each stakeholder are provided in Eq. 3.2 to Eq. 3.6. In our experiments, we focus on comparing two configurations:

- *Baseline Controller:* The ego vehicle operates using a standard baseline controller without the human reasons-based supervision (Rahmani et al., 2023).
- *Baseline Controller with Replanner:* The baseline controller is augmented with the human reasons-based supervision framework, which triggers replanning when the vehicle's behaviour does not align with the predefined ethical thresholds.

These experiments are designed to assess how integrating human reasons-based supervision framework impacts decision-making in vehicle behaviour during ethically challenging situations; specifically in the context of safely overtaking a cyclist on a bidirectional road. To calibrate our reason models, we set the threshold values summarised in Table 3.1 based on empirical studies of cyclist and driver behaviour (Oskina et al., 2023; Hagemeister & Bertram, 2024).

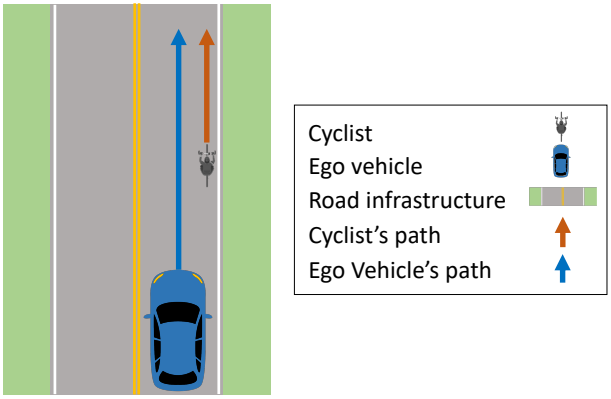


Figure 3.2: Illustration of an ego vehicle approaching a cyclist on a narrow bidirectional road, highlighting the ethical challenge in overtaking due to road constraints.

Table 3.1: Parameter Values for Reason Models

Parameter	Value
$d_{\text{th,vru}}$ (Cyclist's perceived too-close distance threshold)	8 m
$t_{\text{th,vru}}$ (Max time cyclist tolerates close following)	5 s
$d_{\text{th,driver}}$ (Driver's perceived too-close distance threshold)	12 m
$t_{\text{th,driver}}$ (Max time driver tolerates close following)	10 s
τ_{s_i} (Reason alignment threshold for all stakeholders)	0.7

3.5 Results

The results of validation for the controller with and without the proposed human reasons-based supervision framework are depicted in Fig. 3.3 and Fig. 3.4, respectively. The timestamps next to the ego vehicle and the cyclist indicate their positions at that time, helping to visualize their relative movements. The results for the *Baseline controller* suggest that while the system successfully handles basic path planning and collision avoidance, it does not adequately account for human reasons, particularly in terms of the driver's and cyclist's perspectives. As shown in Fig. 3.3.a, the ego vehicle follows the cyclist and reaches the goal in 35 seconds without attempting to overtake. This behaviour demonstrates a stop-and-go dynamic, which is further illustrated in Fig. 3.3.c, where the speed of the ego vehicle is depicted. Initially, the global planner generates a smooth path toward the goal. However, when the ego vehicle approaches the cyclist and a potential collision risk arises, the controller activates a collision avoidance strategy, reducing the ego vehicle's speed to avoid the potential collision. As the distance between the two increases, the controller allows the vehicle to accelerate and realign with the planned path.

Despite effectively tracking the planned trajectory (Fig. 3.3.d), the system's performance in aligning with human reasons decreases over time (Fig. 3.3.b). It is apparent that from the 10th second until the end of the simulation, the driver's reason score for time efficiency decreases sharply to zero by the end of the simulation. This is because the driver must remain patient to stay in the mode of following the vehicle from behind. On the other hand, the cyclist's reason score for comfort fluctuates (due to the fluctuations of the vehicle's speed and distance to the cyclist) but ultimately forms a decay pattern. Over time, the score decreases further due to the accumulation of time spent being followed by the ego vehicle at a close distance. Meanwhile,

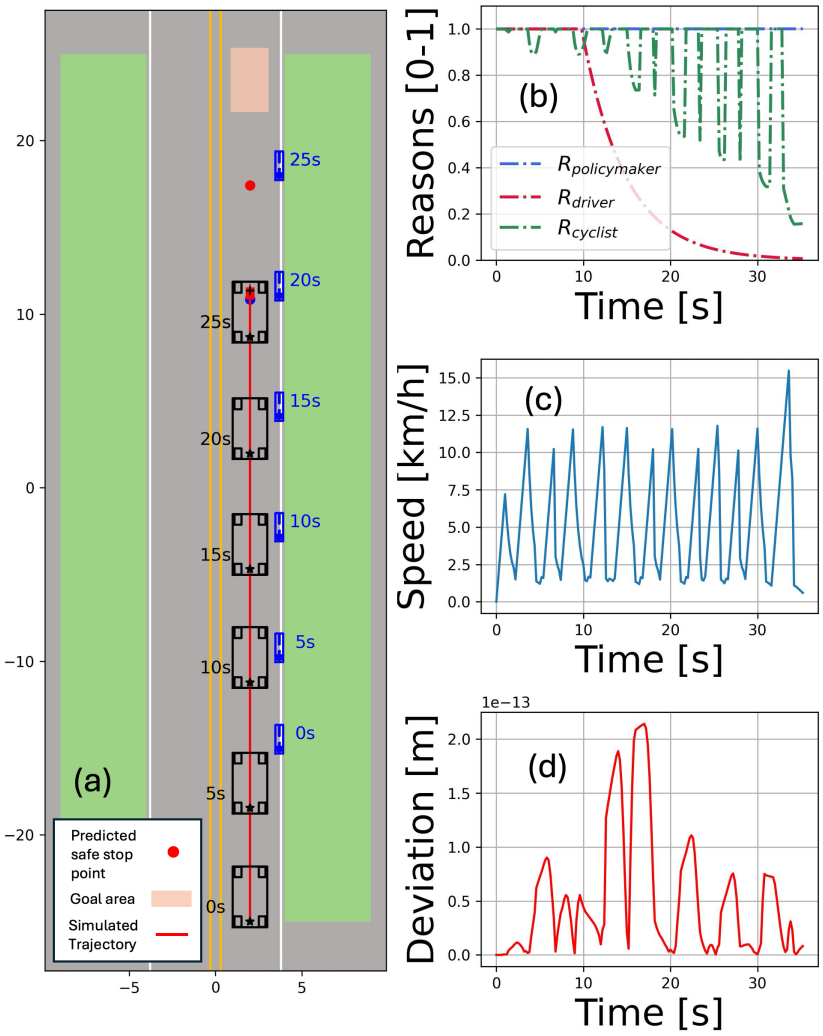


Figure 3.3: Results of running the model in baseline controller

the system's performance demonstrates strong alignment with the road policymaker's reason score for regulatory compliance since the ego vehicle consistently stays in the right lane.

The *Baseline controller with a replanner* allows the ego vehicle to successfully overtake the cyclist, addressing human reason priorities but at the cost of temporary regulatory compliance violations. Initially, the ego vehicle follows the cyclist from behind for the first 11 seconds, adhering to its straight path trajectory while exhibiting stop-and-go behaviour, as shown in Fig. 3.4.c. At the 11.5-second mark, the human reason-based supervision framework detects that the driver's reason score for time efficiency has fallen below its threshold of 0.7, due to the accumulation of the waiting time of the driver and the cyclist. This triggers the planner to generate a new feasible trajectory. This new path briefly crosses the bidirectional road before returning to the right lane to reach the goal. During the overtaking manoeuvre, the close proximity between the ego vehicle and the cyclist causes a temporary decrease in reason scores, and the violation of the right-lane regulation further reduces the policymaker's reason score. However, once the ego vehicle successfully overtakes the cyclist and returns to the intended lane, all reason values recover to one. In this scenario, the ego vehicle achieves the goal in just 18 seconds by overtaking the cyclist, significantly reducing the driver's waiting time. It is worth noting that the

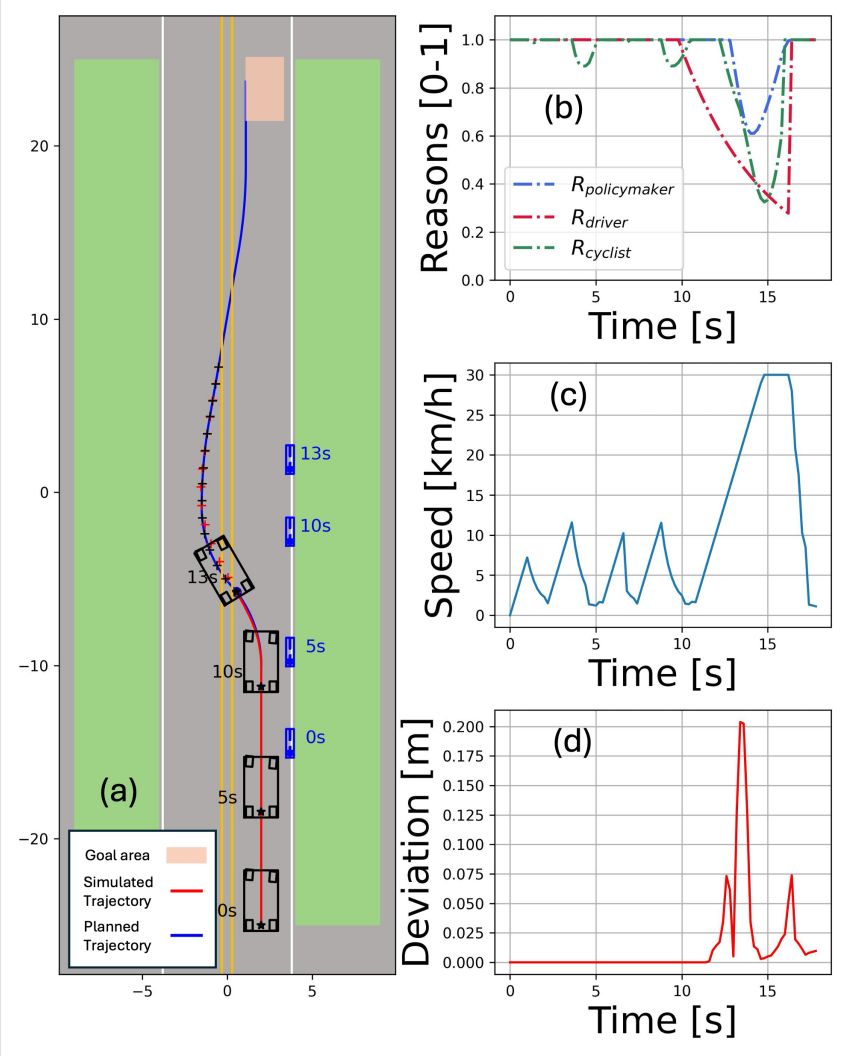


Figure 3.4: Results of running the model in baseline controller with replanner

deviations on the order of centimeters in both scenarios may be attributed to the limitations and constraints of the MPC.

3.6 Discussion

This study reveals three key contributions of the human reasons-based supervision framework: (1) the ability to detect when current system behaviour no longer aligns with the priorities of human stakeholders by monitoring reason score againts adaptive thresholds, (2) the modularity of the framework to adapt controller behaviour without majorly modifying core components like the global planner or MPC settings, and (3) the inherent explainability of decision-making processes, enabling autonomous systems to justify behavioural changes based on stakeholder reason alignment. These features are essential for building trust and acceptance in automated vehicle deployment. The experimental results yield several insights. While both the *Baseline Controller* and the *Baseline Controller with Replanner* successfully guide the ego vehicle to follow the planned trajectory and reach the goal area, they differ significantly in how they

respond to human reasons. This leads to distinct trade-offs in performance and alignment with the priorities of human reasons.

The *Baseline Controller* demonstrates a conservative approach, prioritizing the policymaker's reason for regulatory compliance. This results in strict adherence to rules, but at the cost of neglecting other human reasons priorities. While the controller achieves basic objectives like collision avoidance and goal attainment, it fails to address the driver's reason for time efficiency and the cyclist's reason for comfort.

In contrast, the *Baseline Controller with Replanner* introduces a dynamic and adaptive approach through the human reasons-based supervision component of the framework. Our results show that responding to triggers can lead to decisions that better reflect a balance of human reasons. For example, the system may temporarily violate regulatory compliance (lowering the policymaker's reason score) to reduce discomfort for the cyclist or impatience from the driver. While the trigger does not resolve value conflicts, it signals when the current trajectory may no longer align with a stakeholder's reasons. However, how the planner could systematically select among alternatives is beyond the scope of this research. Future work is needed to extend this supervision layer with decision-making mechanisms that actively weigh and decide how to respond when human reasons are in conflict.

The choice of threshold values plays a critical role in this framework. A lower threshold might delay intervention, leading to prolonged misalignment with human reasons, while a higher threshold could result in overly frequent replanning, which increases the computational cost. Additionally, the vehicle's state when the reasons falls below the threshold—such as its proximity to the cyclist, speed, or surrounding environment—can influence the feasibility of the replanned trajectory. These factors highlight the importance of carefully pick the right threshold to ensure feasibility and stability.

However, while the threshold offers an interpretable mechanism for initiating replanning, it does not capture the full nuance of how human drivers make context dependent trade-offs, such as deciding when it is safe to pass with oncoming traffic or assessing visibility in hilly terrain. Rather than prescribing the best course of action, the framework uses thresholds to signal that the current plan may no longer reflect certain stakeholder priorities. Grounding threshold values in empirical studies, and learning or tuning them from human data, is a promising direction for future work.

Nonetheless, we emphasise that our human reasons-based supervision framework, which triggers replanning based on threshold values, is not intended to compete with existing nuanced decision-making algorithms for dynamic environments, such as multi-policy decision-making (MPDM) proposed by Cunningham et al. (2015); Mehta et al. (2016), but to complement them. While our framework has lower resolution than MPDM's continuous policy evaluation, its strength lies in simplicity. It avoids the computational cost of constant replanning by activating only when a misalignment with human reasons is detected.

Future work could integrate MPDM-style approaches into our architecture, enabling motion planners that not only generate but also evaluate candidate trajectories based on human reasons rather than predefined policies. This integration could support more nuanced balancing of stakeholder priorities while preserving interpretability.

Overall, by aligning system behaviour with human priorities, the proposed framework enables more human-centric decision-making, which is essential for user trust and acceptance in

real-world applications. Thanks to the framework's modularity, its integration into existing automated system architectures is straightforward. Its implementation can be extended to more complex environments, such as urban driving or multi-agent systems, where balancing multiple reasons priorities is critical. Future work could explore enhancing the framework's capabilities, such as using it not only to trigger replanning but also to identify trajectories that maximise human reasons across all agents. Testing in dynamic and unpredictable environments would further validate its robustness and scalability.

This study underscores the importance of incorporating human reason into automated driving systems. The findings demonstrate that while strict regulatory compliance ensures safety and rule-following, mechanisms that detect misalignment with human priorities and prompt reconsideration can lead to decisions that better reflect the reasons of multiple stakeholders. This insight paves the way for future developments in automated systems that are both technically robust and socially and ethically aligned with human reasons.

3.7 Conclusion

This study proposes a human based-reason supervision framework to support automated vehicles (AVs) to navigate routine yet ethically challenging scenarios. The framework introduces a novel approach to AV planning by evaluating whether the vehicle's behaviour aligns with human reason and triggering a replan assignment if misalignment is detected. The key contributions demonstrated through this work are: (1) A detection mechanism for identifying misalignment between AV behaviour and stakeholder reasons based on reason score thresholds; (2) Modular integration into the AV control architecture without modification of the core planner or motion controller; (3) Explainability through the use of stakeholder reason scores, enabling interpretable justifications for behavioural changes. These features enable AVs to align with human reasons in real time, ensuring more human-centric decision-making.

Chapter 4

Trajectory Scoring and Selection

This chapter presents a trajectory scoring and selection framework that evaluates how automated vehicle (AV) decisions align with the reasons of relevant human agents. Building on the reasons-based supervision framework introduced in Chapter 3, it focuses on the evaluative aspect of tracking by examining whether AV trajectory choices reflect the priorities of drivers, vulnerable road users, and policymakers. The chapter introduces a quantitative evaluation method that assigns scores to candidate trajectories based on their alignment with each agent's reasons. The framework incorporates an agent-balance function to discourage neglecting any stakeholder and to promote fairness in decision-making. Simulation studies demonstrate how varying the prioritisation of agents' reasons influences trajectory selection and highlight key trade-offs between rule adherence and safety.

Whereas Chapter 3 focuses on behavioural adjustment through supervision, this chapter addresses reason-based evaluation and selection among candidate trajectories.

This chapter is based on the following paper, accepted for publication at the 2025 5th International Conference on Robotics, Automation, and Artificial Intelligence (RAAI):

Suryana, L. E., Rahmani, S., Calvert, S. C., Zgonnikov, A., and van Arem, B. (2026). A Framework for Human-Reason-Based Trajectory Evaluation in Automated Vehicles. In Proceedings of the 5th RAAI (pp. 734–741). IEEE.

4.1 Abstract

One major challenge for the adoption and acceptance of automated vehicles (AVs) is ensuring that they can make sound decisions in everyday situations that involve ethical tension. Much attention has focused on rare, high-stakes dilemmas such as trolley problems. Yet similar conflicts arise in routine driving when human considerations, such as legality, efficiency, and comfort, come into conflict. Current AV planning systems typically rely on rigid rules, which struggle to balance these competing considerations and often lead to behaviour that misaligns with human expectations. This paper introduces a reasons-based trajectory evaluation framework that operationalises the tracking condition of Meaningful Human Control (MHC). The framework represents human agents' reasons (e.g., regulatory compliance) as quantifiable functions and evaluates how well candidate trajectories align with them. It assigns adjustable weights to agent priorities and includes a balance function to discourage excluding any agent. To demonstrate the approach, we use a real-world-inspired overtaking scenario, which highlights tensions between compliance, efficiency, and comfort. Our results show that different trajectories emerge as preferable depending on how agents' reasons are weighted, and small shifts in priorities can lead to discrete changes in the selected action. This demonstrates that everyday ethical decisions in AV driving are highly sensitive to the weights assigned to the reasons of different human agents.

4.2 Introduction

Evaluating how automated vehicles (AVs) handle ethically challenging situations in everyday driving is essential for their adoption and acceptance by society (Lin, 2016; Millar et al., 2017). Such situations often require trade-offs between competing values, such as safety, legality, and social norms, for which no clear or universally optimal solution exists. For instance, an AV may need to decide whether to cross a solid line to safely overtake a cyclist (FSDEvolution, 2025) or whether to come to a full stop at an empty junction when no vehicles or pedestrians are present (Tesla, 2025). While human drivers usually make such choices intuitively, AVs face greater difficulty because they often depend on rule-based systems or predefined optimisation algorithms (Aksjonov & Kyrki, 2021; Yuan et al., 2024). These systems struggle to balance safety, efficiency, regulatory compliance, and social expectations in real time, which can result in decisions that diverge from human judgement and values (Bin-Nun et al., 2022).

Addressing these dilemmas remains largely unaddressed in current AV design paradigms (Himmelreich, 2018). Most existing approaches to ethical decision-making focus on rare, extreme situations, such as the well-known "trolley problem" (Bonnefon et al., 2019). While such scenarios are philosophically intriguing, they are seldom encountered in routine driving. As Lin observes (Lin, 2016), everyday ethical challenges go well beyond rare, binary dilemmas. They require flexible, context-aware reasoning—something current AV algorithms often struggle to achieve. Similarly, Nyholm nyholm2016ethics argues that focusing too heavily on extreme scenarios oversimplifies the probabilistic and dynamic nature of real-world driving environments.

Addressing day-to-day ethical challenges requires reasoning that considers the diverse goals of multiple human agents. In this research, we use the term human agents to include not only direct road users, such as drivers, cyclists, and pedestrians, but also those indirectly affected,

including policymakers and society (Calvert & Mecacci, 2020). These agents may prioritise safety, legality, efficiency, or social norms differently. As a result, ethical tensions arise when AVs must navigate between competing expectations. Recent work (Cecchini et al., 2024; Henschke, 2020) has called for more holistic approaches. Such approaches aim to integrate deontological, consequentialist, and virtue-based principles while also ensuring transparency and alignment with human moral intuitions.

However, integrating ethical principles into AV decision-making remains a challenge. Recent approaches have proposed ethical trajectory planning algorithms grounded in deontological reasoning (Thornton et al., 2016), or based on risk and cost functions that combine multiple ethical considerations (Geisslinger et al., 2023). While these models represent progress, they have also been critiqued for lacking transparency (Kirchmair & Paulo, 2023). Key concerns include how ethical principles are selected, how conflicts are resolved, and how resulting decisions align with legal and societal expectations.

The principle of Meaningful Human Control (MHC) (Santoni de Sio & Van den Hoven, 2018; Mecacci & Santoni de Sio, 2020) offers a promising foundation to address these critiques. MHC is a design principle with two aims. First, AV behaviour should reflect the intentions and moral reasons of relevant human agents, a requirement known as tracking. Second, it should remain possible to assign responsibility to informed and accountable individuals, a requirement known as tracing (de Sio et al., 2023). To fulfil the tracking condition, AV behaviour must be responsive to the reasons of relevant agents, including their values, plans, and intentions. It must also account for those indirectly affected, such as vulnerable road users and policymakers (Mecacci & Santoni de Sio, 2020).

MHC could serve as a conceptual bridge between ethical principles and observable AV behaviour. It links abstract moral values, such as those in deontological or utilitarian ethics, to practical elements like plans, intentions, and actions (Mecacci & Santoni de Sio, 2020). In this way, MHC suggests that moral values should be reflected in agents' practical choices, including priorities such as safety, comfort, and rule compliance. However, before these principles can guide design, we must first evaluate whether AV decisions actually reflect them. Without a systematic evaluation method, it is impossible to judge whether an AV's behaviour aligns with ethical expectations such as fairness, harm minimisation, or accountability. Although recent work has helped clarify the concept of MHC, the challenge remains: how can it be applied in practice to evaluate AV behaviour? This motivates the need for a framework capable of assessing whether AVs act in accordance with the moral reasons of relevant human agents.

To address this need, we propose a reason-based evaluation framework that measures how well planned AV trajectories align with the reasons of relevant human agents. Beyond trajectory alignment, the framework also tests whether an AV system satisfies the tracking condition of MHC in practice. It follows the evaluation procedure outlined by Suryana et al. (2024), which involves three steps: identifying relevant agents and their reasons, specifying the AV behaviours that should reflect those reasons, and conducting the reason evaluation. Our framework does not replace existing trajectory planning methods but evaluates their outcomes. In doing so, it provides a transparent way to determine whether a chosen trajectory aligns with the moral reasons of the agents involved.

To illustrate our approach, consider a scenario where an AV follows a slow cyclist on a road marked with double solid yellow lines, which prohibit overtaking (FSDEvolution, 2025). After

a few seconds, a human driver intervenes and overtakes, exposing a misalignment between the AV's rule-based behaviour and human judgement. Our framework evaluates such cases by modelling the priorities of relevant agents, such as policymakers, vulnerable road users, and passengers, as mathematical functions. These functions are then used to score and compare candidate trajectories, similar to existing motion-planning pipelines. The key difference is that, instead of optimising for fixed performance criteria, we assess alignment with human reasons, providing a new layer of ethical evaluation.

Specifically, this paper introduces a novel approach for evaluating whether AV behaviour in everyday ethically challenging scenarios reflects the reasons of relevant human agents. Our primary contributions are:

1. We develop a **reasons-based trajectory evaluation framework** that measures the alignment between AV trajectories and the reasons of relevant human agents. This allows us to assess whether the system satisfies the tracking condition of MHC in practice.

2. We demonstrate, through simulation, that the framework supports ethically grounded and interpretable decision-making. It does so by modelling agent influence as both quantifiable and adjustable, and by enabling both forward and inverse analysis of decisions.

The remainder of this paper is organised as follows: Section 4.3 presents the methodology. Section 4.4 describes the experimental setup. Sections 4.5 and 4.6 report and discuss the results. Section 4.7 concludes the paper.

4.3 Methodology

Current AV decision-making systems lack a mechanism to evaluate whether a selected trajectory aligns with the reasons of agents affected by it. To address this, we propose a unified trajectory scoring function that integrates agent importance, reason-level evaluations, and a fairness adjustment, thereby supporting the tracking condition of Meaningful Human Control (MHC). We begin by defining the components of the framework, then build up to the final scoring formulation.

4.3.1 Human Agents and Their Reasons

We define the human agent set $H = \{h_1, h_2, \ldots, h_n\}$, where each human h_i has a set of reasons

$$\mathcal{R}_i = \{r_{i1}, r_{i2}, \ldots, r_{im_i}\},$$

with m_i denoting the number of reasons associated with human h_i. Here, $b \in \{1, \ldots, m_i\}$ indexes the individual reasons of human h_i. Each human is assigned a weight $w_i \in [0,1]$, with $\sum_{i=1}^{n} w_i = 1$. Each agent aggregates their reasons using weights $\alpha_{ib} \in [0,1]$, where $\sum_b \alpha_{ib} = 1$.

4.3.2 Trajectories and Environment Representation

Given candidate trajectories $T = \{T_1, \ldots, T_k\}$, each $T_a \in T$ is a discretized sequence of ego states:

$$T_a = \{s_{a0}, s_{a1}, \ldots, s_{ap}\},$$

where $a \in \{1, \ldots, k\}$ indexes candidate trajectories, p denotes the number of discrete time steps, and s_{al} is the ego vehicle's state at time step $l \in \{0, \ldots, p\}$. Each time step corresponds to $t_l = l \cdot \Delta t$, where Δt is the planning time resolution. States include position, orientation, velocity, and other kinematic quantities, and are generated via feasible motion models.

In real driving, the ego vehicle's trajectory must account for dynamic entities such as other vehicles, pedestrians, and cyclists. Since these entities influence whether human reasons can be fulfilled (e.g., safety or comfort), we include their trajectories in the evaluation.

Dynamic entities are indexed by $q \in \{1, \ldots, Q\}$, with trajectories

$$E_q = \{e_{q0}, \ldots, e_{qp}\},$$

and we denote the set of all such trajectories as $\mathcal{E} = \{E_1, \ldots, E_Q\}$. At time t_l, the environment snapshot is

$$\mathcal{E}_l = \{e_{ql} \mid q = 1, \ldots, Q\}.$$

4.3.3 Reason-Level Evaluation

Each reason r_{ib} has a per-time-step evaluation function

$$f_{ib}(s_{al}, \mathcal{E}_l, t_l) : (s_{al}, \mathcal{E}_l, t_l) \to [0, 1],$$

and a trajectory-level score obtained via a temporal aggregation operator:

$$F_{ib}(T_a, \mathcal{E}) = \Phi\left(\{f_{ib}(s_{al}, \mathcal{E}_l, t_l)\}_{l=0}^{p}\right). \tag{4.1}$$

Here, Φ maps per-time-step evaluations to a trajectory-level value. In this work, we adopt the uniform average

$$\Phi = \frac{1}{p+1} \sum_{l=0}^{p} (\cdot), \tag{4.2}$$

though alternative operators (e.g., weighted multi-objective sums (Xu et al., 2012) or cumulative integral costs (Williams et al., 2017)) may be used depending on design requirements or normative assumptions.

4.3.4 Aggregating Reasons and Agents

Each agent's reason-level score is

$$S_i(T_a) = \sum_{b=1}^{m_i} \alpha_{ib} F_{ib}(T_a, \mathcal{E}). \tag{4.3}$$

Combining these across agents gives the unbalanced score:

$$S_w(T_a) = \sum_{i=1}^{n} w_i S_i(T_a). \tag{4.4}$$

4.3.5 Agent Balance Function

To ensure equitable agent influence and preserve MHC, we introduce an agent balance function $B(\mathbf{w}, \mathbf{w}^*)$, which penalizes highly skewed weight configurations:

$$B(\mathbf{w}, \mathbf{w}^*) = \left(1 - \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(w_i - w_i^*)^2}}{\sqrt{\sum_{i=1}^{n}(w_i^*)^2}}\right) \cdot \min_i \left(\frac{w_i}{w_i^*}\right) \tag{4.5}$$

where $\mathbf{w}^*$ is the ideal distribution, typically uniform ($w_i^* = 1/n$). The first term measures deviation from the ideal via RMS error, while the second ensures no agent is excluded (i.e., $w_i > 0$). Together, they promote proportional fairness and representation, addressing concerns in (Calvert et al., 2020b; Mecacci & Santoni de Sio, 2020) about agent exclusion in autonomous systems.

4.3.6 Final Scoring and Trajectory Selection

The final balanced score is

$$S(T_a) = B(\mathbf{w}, \mathbf{w}^*) \cdot S_w(T_a). \tag{4.6}$$

After computing $S(T_a)$ for all trajectories, we select the one that maximises alignment with human reasons:

$$T^* = \arg\max_{T_a \in T} S(T_a). \tag{4.7}$$

In this work, this selection is used solely for evaluation and comparison purposes, illustrating which trajectory best aligns with human reasons.

Figure 4.1, adapted from the framework structure in (Suryana et al., 2025d), illustrates how the proposed human-reasons-based trajectory evaluation module integrates into a standard hierarchical AV decision-making stack. In a conventional architecture, the global planner typically generates a single nominal trajectory that is passed directly to the local controller. Following the practice in (Rahmani et al., 2025), we instead assume that the global planner can generate a set of feasible candidate trajectories and select among them according to specific evaluation criteria. In our framework, this set of candidate trajectories is intercepted before reaching the controller and evaluated by the human-reasons module. After scoring, the global planner selects the trajectory with the highest alignment score, T^*, and returns it to the standard control pipeline for execution.

In this way, our method does not alter the control architecture itself; rather, it inserts a normative evaluation layer that ensures Meaningful Human Control over the trajectory-selection stage.

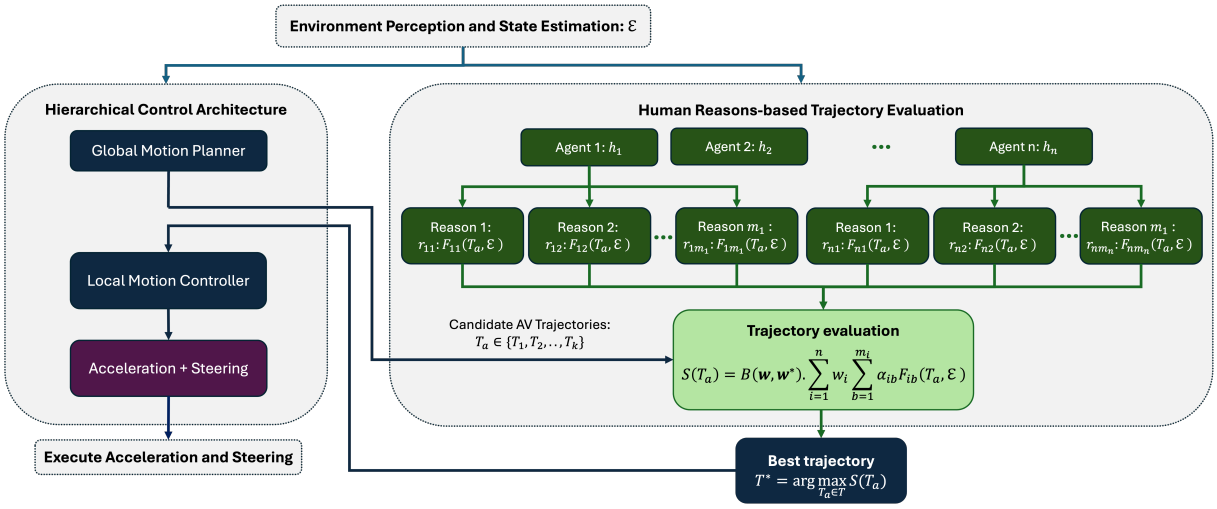


Figure 4.1: Integration of the proposed human-reasons-based trajectory evaluation module into a hierarchical AV control architecture. The module does not generate or select trajectories; instead, it evaluates candidate trajectories produced by the global planner for alignment with human reasons. The global planner then uses these scores to select the trajectory that best satisfies both motion-planning and human-reason considerations.

4.4 Experimental Setup

4.4.1 Overtaking Scenario Description

To demonstrate our reasons-based trajectory evaluation framework, we implement an ethically challenging overtaking scenario involving three agents: a policymaker, a driver, and a cyclist. The scenario is adapted from a real-world case FSDEvolution (2025), where Tesla's Full Self-Driving Beta chose to remain behind a cyclist on a no-passing road, while a human driver ahead illegally overtook—highlighting tensions between safety, legality, and efficiency.

This situation reflects conflicts between regulatory compliance (policymaker), travel efficiency (driver), and safety/comfort (cyclist). The AV must decide whether to stay behind or overtake, trading off compliance for potential gains in efficiency. The AV encounters a slow-moving cyclist (5 km/h) on a rural two-lane road (7 m wide, 3.5 m per lane) with no oncoming traffic and a 30 km/h speed limit. A visual depiction, including the AV's trajectories, is shown in Fig. 4.2.

4.4.2 Agents and Their Reasons

Suryana et al. (2024) evaluated safety reason alignment in partially automated driving systems using a simplified setting with two human agents and a single shared reason. While their study introduced a foundational approach to reason-based evaluation, it did not address conflicts that may arise between distinct agents with differing priorities.

To explore such conflicts, this work models three agents, each associated with their own reason. These agents reflect a range of viewpoints commonly encountered in AV scenarios.

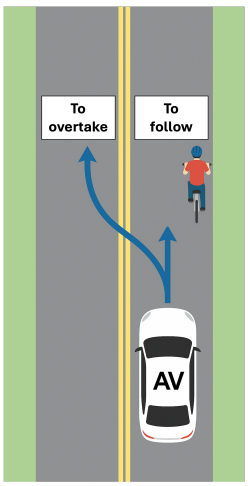


Figure 4.2: Illustration of the vehicle-cyclist overtaking scenario showing the initial configuration, possible trajectories, and relevant parameters.

While we focus on three agents for illustration, the framework can scale to any number of human agents, as each agent is represented as a vector $w \in \mathbb{R}^n$.

The **policymaker** (h_1) prioritises regulatory compliance, such as maintaining lane discipline and ensuring the vehicle returns to the correct lane after overtaking. The **driver** (h_2) values time efficiency, aiming to minimise delays caused by slower vehicles while still maintaining safety. Meanwhile, the **cyclist** (h_3) is concerned with safety and comfort, which includes maintaining sufficient lateral clearance and expecting appropriate overtaking behaviour from surrounding vehicles. Each agent uses a single reason ($\alpha_{i1} = 1$) with equal initial weight ($w_i = 1/3$). We explore other weight configurations in a sensitivity analysis.

4.4.3 Candidate Trajectories

We define four candidate AV trajectories $\mathcal{T} = \{T_1, T_2, T_3, T_4\}$, representing different patterns of agent prioritisation in the overtaking scenario. These trajectories vary in clearance distance, lane use, and alignment with the reasons of drivers, cyclists, and policymakers. Their generation follows the procedure outlined by Rahmani et al. (2023), which provides a structured approach for producing AV trajectories in interaction with surrounding agents. To generate these four alternatives, we experimented with the heuristic function in the global planner introduced by Rahmani et al. (2023); however, the details of this adaptation are beyond the scope of this paper.[1]

Rather than presenting a binary decision, such as death or alive, this setup reflects the kind of everyday ethical challenges AVs are more likely to encounter—such as balancing safety, legality, and mobility. This design aligns with the critique of trolley problem framings offered by Himmelreich (2018), who advocate for a shift towards mundane driving scenarios that require context-sensitive reasoning rather than abstract moral binaries.

Figure 4.3 illustrates the four candidate trajectories considered in this study. The ego vehicle is depicted as a black box, and the cyclist as a blue box. The solid blue line represents the

[1] Trajectory generation code: https://github.com/lucassuryana/AV-Simulation

cyclist's trajectory, assumed to continue forward at a constant speed. The four ego trajectories ($T_1 - T_4$) differ in their lateral clearance and overtaking behaviour:

- Trajectory 1 (T_1): Small-gap overtaking — minimal clearance; prioritises driver convenience, with limited regard for cyclist and policymaker considerations.
- Trajectory 2 (T_2): Medium-gap overtaking — balanced clearance, moderate consideration of cyclist and policymaker.
- Trajectory 3 (T_3): Large-gap overtaking — wide clearance, prioritises cyclist comfort and safety, least compliant with policymaker expectations.
- Trajectory 4 (T_4): Conservative following — no overtake, prioritises legal compliance, lowest consideration for driver efficiency.

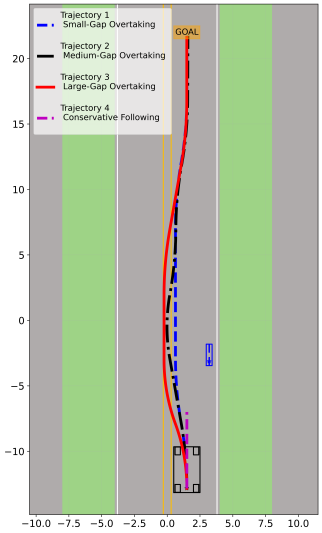


Figure 4.3: Spatial visualisation of four candidate AV trajectories (T_1–T_4) relative to a cyclist. The trajectories vary in lateral clearance and lane usage, reflecting different prioritisation patterns across safety, efficiency, and legal compliance.

4.4.4 Evaluation Functions and Implementation

Each agent's evaluation is computed via a per-time-step function $f_{ib}(s_{al}, \mathcal{E}_l, t_l)$, introduced in Section 4.3, and averaged over the trajectory duration (Equation 4.1). In our scenario, each agent as described in Section **??** has a single reason ($b = 1$ and $\alpha_{ib} = 1$); therefore, we denote reason-evaluation functions as f_1, f_2, and f_3 for readability.

Policymaker Evaluation Focusing on lane compliance, the policymaker's evaluation is:

$$f_1(s_{al}, \mathcal{E}_l, t_l) = \begin{cases} 1, & d_{\text{veh}}(s_{al}) > 0, \\ e^{k_1 \cdot d_{\text{veh}}(s_{al})}, & \text{otherwise}, \end{cases} \tag{4.8}$$

where $d_{\text{veh}}(s_{al})$ is the lateral distance from the lane centerline, and $k_1 = 0.2$ controls penalty severity.

Driver Evaluation To model time efficiency, we define a cumulative follow time $t_{\text{elapsed},l}$ (initialized as 0). It updates each step by Δt if the AV is within d_{driver} of a cyclist: $t_{\text{elapsed},l+1} = t_{\text{elapsed},l} + \Delta t$ if $d_{\text{vc}} \leq d_{\text{driver}}$, else unchanged.

The driver's evaluation is:

$$f_2(s_{al}, \mathcal{E}_l, t_l) = \begin{cases} 1, & t_{\text{elapsed},l} < t_{\text{driver}} \\ & \quad \vee\, d_{\text{vc}} > d_{\text{driver}}, \\ \frac{1}{e^{k_2(t_{\text{elapsed},l} - t_{\text{driver}})}}, & \text{otherwise}, \end{cases} \tag{4.9}$$

where d_{vc} is the distance to the cyclist, and $k_2 = 0.2$. This formulation is supported by behavioural studies showing that driver patience declines with prolonged close following. Naveteur et al. (2013) link waiting time and time pressure to rising impatience. Together, these findings justify modeling satisfaction as a decaying function of follow time.

Cyclist Evaluation The cyclist's evaluation combines spatial safety and temporal comfort:

$$f_3(s_{al}, \mathcal{E}_l, t_l) = R_{sa}(s_{al}, \mathcal{E}_l) \cdot R_{cp}(s_{al}, \mathcal{E}_l, t_{\text{follow},l}) \tag{4.10}$$

Spatial safety component:

$$R_{sa}(s_{al}, \mathcal{E}_l) = \begin{cases} 1, & d_{\text{vc}} > d_{\text{th}}, \\ \frac{1}{e^{k_3(d_{\text{th}} - d_{\text{vc}})}}, & \text{otherwise}, \end{cases} \tag{4.11}$$

Spatial temporal comfort component: The follow time $t_{\text{follow},l}$ (initially 0) updates as $t_{\text{follow},l+1} = t_{\text{follow},l} + \Delta t$ if $d_{\text{vc}} \leq d_{\text{th}}$, else unchanged. The comfort score is:

$$R_{cp}(s_{al}, \mathcal{E}_l, t_{\text{follow},l}) = \begin{cases} 1, & t_{\text{follow},l} < t_{\text{th}} \\ & \quad \vee\, d_{\text{vc}} > d_{\text{th}}, \\ \frac{1}{e^{k_4(t_{\text{follow},l} - t_{\text{th}})}}, & \text{otherwise}, \end{cases} \tag{4.12}$$

Constants: $k_3 = k_4 = 0.2$, Δt is the time step, and d_{th}, t_{th} are the cyclist's safety thresholds. This formulation aligns with findings from Oskina et al. (2023), showing that cyclists adapt behaviour—such as increasing speed and reducing lateral spacing—when followed for extended periods, indicating rising discomfort and feeling unsafe.

4.4.5 Balance Function Implementation

As per Section 6.3, the balance function $B(\mathbf{w}, \mathbf{w}^*)$ penalizes uneven agent weightings. For equal weights ($w_i = 1/3$), $B = 1$; for $w_2 = 0.6$, $w_1 = w_3 = 0.2$, we get $B = 0.487$. Fig. 4.4 shows the balance values across the weight simplex. The function peaks with equal influence and reaches 0 when any agent is excluded ($w_i = 0$).

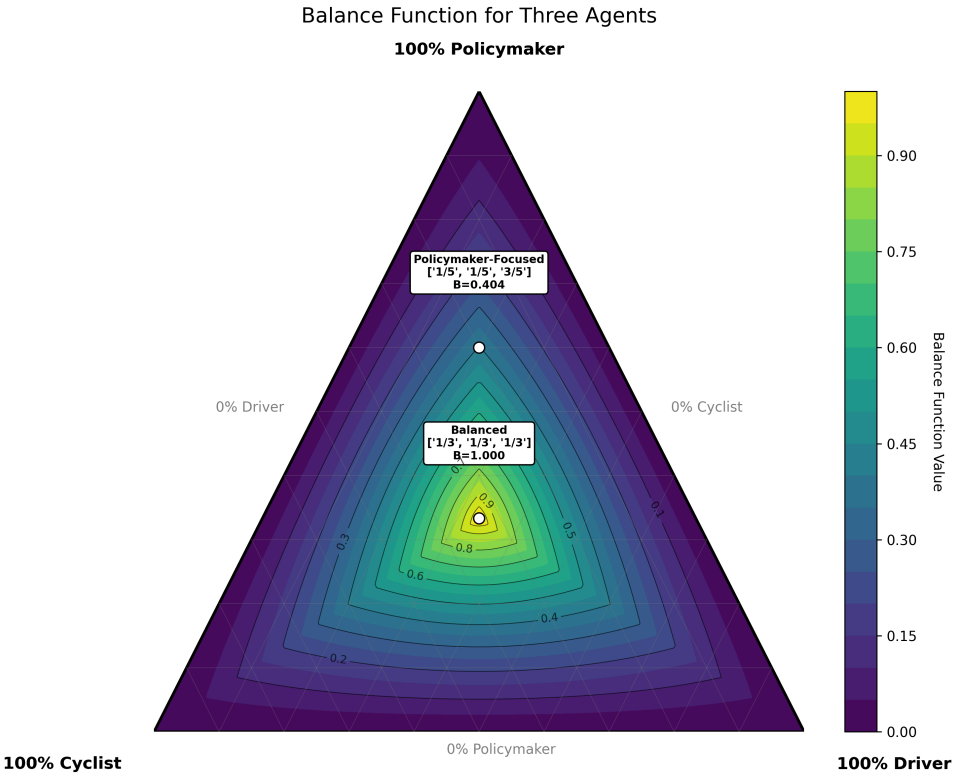


Figure 4.4: Ternary plot showing the output of the balance function $B(\mathbf{w})$ across combinations of agent weights. Maximum balance occurs when all weights are equal.

4.5 Results

This section presents simulation results from the overtaking scenario, where each trajectory was evaluated based on its alignment with agents' reasons. Scores were computed both per agent and in aggregate using equal weighting ($w_i = 1/3$).

We first evaluate alignment under equal agent weighting. Figure 4.5 shows the evaluation results for the four candidate trajectories. The final score $S(T_a)$ quantifies how well each trajectory aligns with the reasons of the policymaker, driver, and cyclist. The red region represents the historical progression of reason-based scores and the triggering condition for supervision, as established in previous work by Suryana et al. (2025d). Once the score drops below the 0.7 threshold, the system generates several alternative trajectories. The blue region then begins—this marks the activation of our reason-based evaluation framework, which re-assesses the new trajectories in terms of alignment with agents' reasons.

Among the four options, Trajectory 1 (Small-Gap Overtake) achieves the highest overall score under equal weighting, while Trajectory 4 (Conservative Following) records the lowest. This suggests that in this context, overtaking with minimal clearance better satisfies the tracking requirement across agents than remaining behind. However, trajectory rankings vary significantly depending on how agents' importance is weighted.

To explore this sensitivity, we varied two agents' weights while keeping the third constant. The resulting trajectory preferences are visualised in the ternary plot in Figure 4.6, illustrating how the optimal choice depends on agent prioritisation.

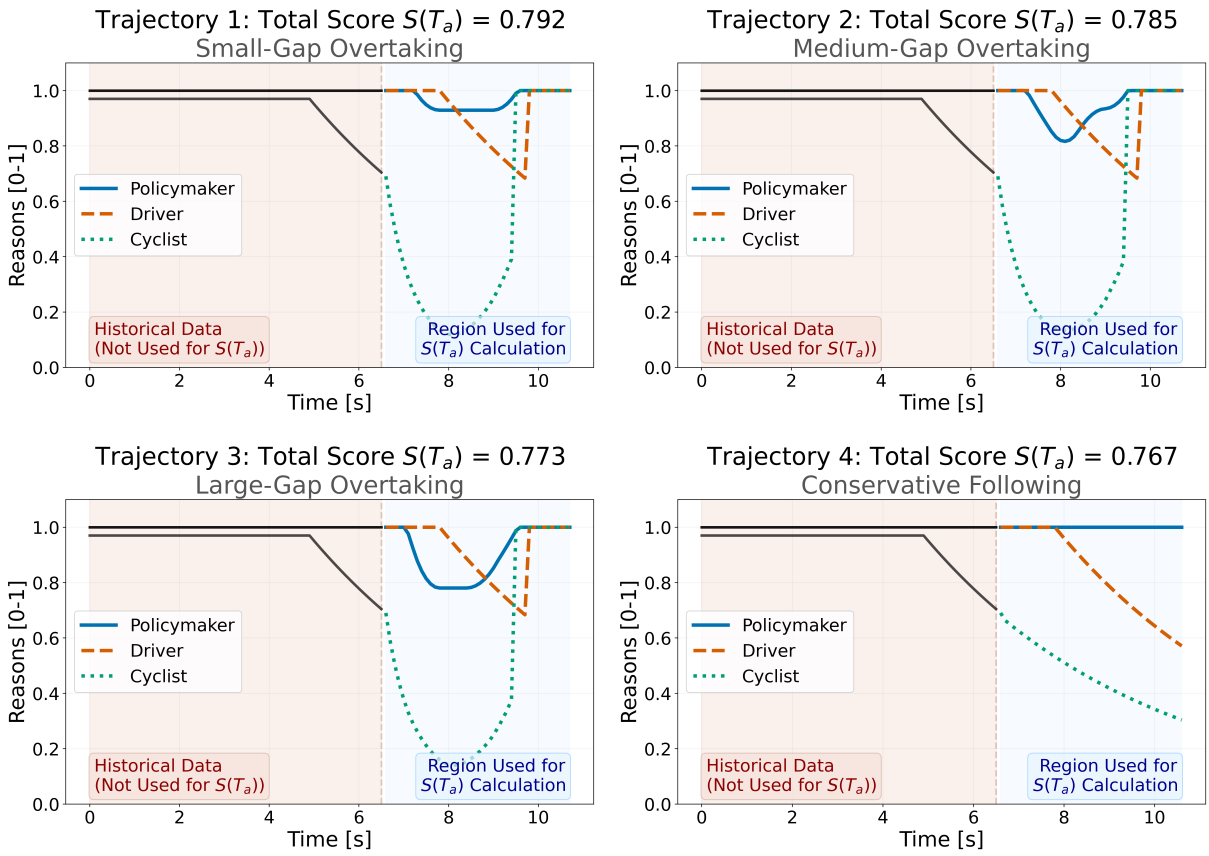


Figure 4.5: Trajectory scores for four candidate trajectories evaluated against agents' reasons. The red region shows historical score progression; the blue region begins when the score drops below 0.7, prompting trajectory reevaluation.

Colored regions indicate which of the four trajectories achieves the highest score under each weight configuration. Blue (Trajectory 1) reflects strong driver prioritisation; Yellow (Trajectory 3) favors the cyclist; and Red (Trajectory 4) aligns with the policymaker. Other colors represent tie cases. Notably, when one agent receives zero weight (along triangle edges), all scores converge, and no clear preference emerges. White contour lines indicate score magnitudes; higher scores concentrate near regions of balanced agent influence.

These results highlight that minor shifts in agents' weights can lead to discrete changes in trajectory preference. Such critical thresholds underscore the ethical sensitivity of AV decision-making and the importance of transparent value prioritisation.

4.6 Discussion

Our reasons-based evaluation framework enables automated vehicles (AVs) to assess candidate trajectories by measuring their alignment with agents' reasons. It assigns weights to each agent, computes scores, and shows how different prioritisations influence decision outcomes.

Scenario illustration and normative tension: The simulation reflects the real-world case described in Section 4.4.1, where strict rule-following created a misalignment between the AV's behaviour and the reasons of relevant agents. Through weighting, the framework captures such

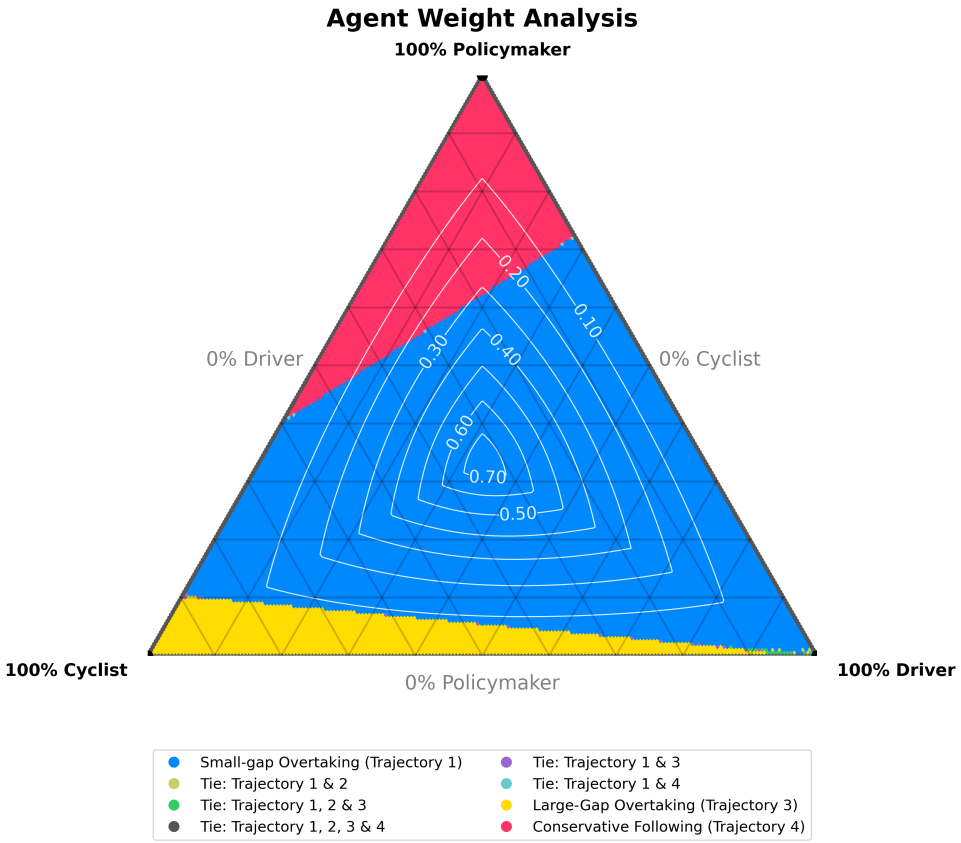


Figure 4.6: Agent Weight Sensitivity: Optimal Trajectory Selection Across Different Priority Distributions

misalignments and demonstrates how adjusted priorities can lead to alternative trajectories. These alternatives may involve short-term trade-offs but achieve closer alignment with agents' collective reasons.

In the overtaking case, the chosen trajectory briefly enters the oncoming lane before returning. Although it achieved the highest aggregate score, it violated traffic rules and conflicted with public expectations of strict AV compliance (Leenes & Lucivero, 2014). Rather than endorsing such violations, the framework highlights the tensions that arise when concerns beyond regulation—such as safety and comfort—are taken into account. A similar dilemma appears when a driver mounts a kerb to let an emergency vehicle pass: technically illegal, yet often seen as serving the common good (Bonnefon et al., 2020).

This example also illustrates how ethical principles surface indirectly through agents' reasons and resulting trajectories. Prioritising safety and comfort reflects consequentialist reasoning, which focuses on outcomes. In contrast, prioritising regulation reflects deontological reasoning, which stresses rule adherence. The framework does not encode these theories directly, but their influence becomes visible through structured reasoning and trajectory evaluation.

Flexibility in prioritisation: The overtaking case highlights one type of tension, but the framework can also accommodate AV designs that prioritise strict regulatory compliance. Giving more weight to policymakers' reasons naturally downplays comfort and efficiency. Yet adjusting weights alone may not always change the decision. In some cases, the balance function $B(\mathbf{w}, \mathbf{w}^*)$ must also be updated so that the evaluation favours regulatory compliance. This

reflects the principle of tracking in Meaningful Human Control.

As shown in the ternary plot, even small weight adjustments can trigger abrupt shifts in the selected trajectory. These threshold effects underline the importance of designing weight-setting strategies carefully and, when needed, updating the balance function to match intended design priorities.

Scalability and modular integration: Beyond single-case illustrations, the framework is modular and can be added to existing AV motion-planning stacks. It operates as an evaluation layer over candidate trajectories, enabling the selection of the option that best balances agents' reasons. Because it works at the evaluation layer, the framework can integrate with both modular pipelines and end-to-end learning-based planners (Teng et al., 2023), without requiring major changes to core control systems.

Transparency and interpretability: A further benefit is interpretability. By quantifying agents' reasons and assigning weights, the framework turns moral values into operational factors that directly influence trajectory selection. For example, if a chosen trajectory scores lower on regulatory compliance but higher on safety and comfort, this trade-off can be surfaced and examined.

Interpretability works in two directions. Forward interpretability checks whether trajectory selection matches predefined agent priorities. Inverse interpretability, by contrast, infers which weight configurations could have produced a given decision. Together, these support transparency by design, as proposed by Felzmann et al. (2020).

This transparency could also benefit regulators. During type approval, for instance, authorities could use the framework to check whether an AV's planned behaviour aligns with ethical expectations such as fairness and accountability European Union (2018). They could do this without access to proprietary source code, since the framework functions as a white-box layer over decision outputs, revealing how behaviours reflect agents' reasons.

Operationalising meaningful human control: In addition to interpretability, the framework supports the tracking condition of Meaningful Human Control. The score function $S(T_a)$ measures how well each trajectory aligns with human reasons, while the balance function $B(\mathbf{w}, \mathbf{w}^*)$ discourages ignoring any agent. By preventing complete exclusion, the framework helps ensure that AV behaviour remains responsive to human reasons.

Limitations and future directions: Several limitations remain. First, the framework currently assumes equal weighting across agents. While this simplifies evaluation, real-world contexts often demand unequal prioritisation—for example, stronger emphasis on safety or regulation. The balance function discourages exclusion but does not prescribe appropriate weight settings or whether they should adapt dynamically. Future work should investigate principled methods for assigning and adjusting weights.

Second, the framework assumes a correct mapping between agents' reasons and their formal representations. This overlooks interpretive challenges in human–AV interaction. For example, regulatory compliance may be modelled as continuous, but some agents (such as law enforcement) may view it as binary. Such mismatches could undermine perceived alignment. Future studies should explore how humans interpret AV actions and whether they feel their reasons are being tracked.

Third, our evaluation focuses on a simplified overtaking scenario involving one AV and one

cyclist. It does not yet capture complex planning problems, such as dense traffic, multi-agent negotiation, or long-term strategies. Extending the framework to richer scenarios would help align it more closely with real-world challenges.

Finally, future work could apply the framework to trajectories generated by different planning systems to compare their ethical alignment. Beyond AVs, the approach could be generalised to other robotic systems that rely on trajectory planning in ethically sensitive situations.

4.7 Conclusion

In this work, we presented a reasons-based trajectory evaluation framework for AVs that supports decisions aligned with human agents' reasons. The framework enables principled comparison of candidate trajectories by quantifying their alignment with agent perspectives and weighting them according to assigned priorities. Our results show that no single trajectory is universally optimal across scenarios. Instead, the best choice depends on how agent weights are configured, with different weighting schemes leading to different outcomes. This highlights the importance of carefully defining agent priorities and examining how these priorities shape AV decision-making. The framework also improves transparency by making the reasoning behind trajectory selection explicit and by supporting validation under the tracking principle of meaningful human control. Although our evaluation is simulation based, the findings demonstrate the framework's value as a tool for assessing how AV decisions reflect agent reasons and provide a foundation for future empirical studies. Further work should investigate how to derive agent weights empirically, test the framework in real-world AV decision-making, and explore its applicability to other robotic systems.

Chapter 5

Perceptions of Safety and Trust in Meaningful Human Control

This chapter examines how the concept of Meaningful Human Control (MHC) relates to drivers' subjective perceptions of safety and trust in partially automated driving systems. While previous chapters focused on the design and computational implementation of human-reason-based frameworks, this chapter initiates the evaluation aspect of the thesis by asking how well the tracking condition of MHC is satisfied in real-world driving experiences. Using qualitative interview data from Tesla "Full Self Driving" (FSD) Beta users, it explores how the system tracks drivers' reasons for action, such as braking and lane changing, and how this ability relates to users' perceived safety and trust. The analysis identifies alignment points between perceived safety, trust, and the degree of MHC, as well as cases where misalignment occurred despite technically safe behaviour. These findings reveal how factors such as reliability, transparency, and ease of driver intervention shape users' experience of control and trust in automation.

Chapters 5 and 6 draw on the same dataset of 103 semi-structured interviews with Tesla FSD Beta and Autopilot users, originally collected to develop a conceptual framework for automation disengagement (Nordhoff, 2024). Both chapters rest on two premises. First, failures of automated systems to track drivers' reasons may underlie drivers' disengagement from the supervisory role. Second, no existing automated driving system has been designed under MHC principles, yet any automated driving system inherently satisfies the tracking condition to some extent (Mecacci & Santoni de Sio, 2020), making real-world systems appropriate for evaluating how far current practice falls from the MHC standard. This chapter focuses on the tracking condition, examining how alignment between system behaviour and drivers' reasons relates to perceived safety and trust. This chapter initiates the evaluation aspect of the thesis by examining MHC from the perspective of user experience rather than system design.

This chapter is based on the following paper, published at the 2024 IEEE Intelligent Vehicles Symposium (IV):
Suryana, L. E., Nordhoff, S., Calvert, S. C., Zgonnikov, A., and van Arem, B. (2024). A Meaningful Human Control Perspective on User Perception of Partially Automated Driving Systems: A Case Study of Tesla Users. In Proceedings of the 2024 IEEE IV (pp. 409–416). IEEE.

5.1 Abstract

The use of partially automated driving systems raises concerns about potential responsibility issues, posing risk to the system safety, acceptance, and adoption of these technologies. The concept of meaningful human control has emerged in response to the responsibility gap problem, requiring the fulfillment of two conditions, tracking and tracing. While this concept has provided important philosophical and design insights on automated driving systems, there is currently little knowledge on how meaningful human control relates to subjective experiences of actual users of these systems. To address this gap, our study aimed to investigate the alignment between the degree of meaningful human control and drivers' perceptions of safety and trust in a real-world partially automated driving system. We utilized previously collected data from interviews with Tesla "Full Self-Driving" (FSD) Beta users, investigating the alignment between the user perception and how well the system was tracking the users' reasons. We found that tracking of users' reasons for driving tasks (such as safe manoeuvres) correlated with perceived safety and trust, albeit with notable exceptions. Surprisingly, failure to track lane changing and braking reasons was not necessarily associated with negative perceptions of safety. However, the failure of the system to track expected manoeuvres in dangerous situations always resulted in low trust and perceived lack of safety. Overall, our analyses highlight alignment points but also possible discrepancies between perceived safety and trust on the one hand, and meaningful human control on the other hand. Our results can help the developers of automated driving technology to design systems under meaningful human control and are perceived as safe and trustworthy.

5.2 Introduction

The increasing use of automated driving systems raises concerns about potential responsibility issues, which can impact safety, adoption, and acceptance of these systems (Nyholm & Smids, 2020; Calvert et al., 2020b). In particular, delegating control to automated systems, either partially or fully, could create responsibility gaps —- situations where no human agent is responsible for the behaviour of the system (Matthias, 2004; Santoni de Sio & Mecacci, 2021). The concept of meaningful human control (MHC) recently gained prominence in addressing the responsibility gap problem (Santoni de Sio & Van den Hoven, 2018; de Sio et al., 2023; Mecacci et al., 2023).

This concept posits that humans, not artificial agents, should remain morally responsible for the behaviour of automated systems (Santoni de Sio & Van den Hoven, 2018). This entails that (partially or fully) automated systems should be designed in such a way that humans interacting with the systems maintain some form of *meaningful* control over the system behaviour, even when not in operational control of the system (Santoni de Sio & Van den Hoven, 2018). Recent work on operationalizing meaningful human control made steps towards specific frameworks and design principles for developing automated driving systems (Heikoop et al., 2019; Calvert et al., 2020a; Cavalcante Siebert et al., 2023). However, there is currently a lack of understanding of how meaningful human control relates to subjective experiences of actual humans, in particular human drivers/users of driving automation.

In this paper, we aim to investigate the alignment between the degree of meaningful human

control and drivers' perceptions of safety and trust in a real-world partially automated driving system. To this end, analyzing subjective evaluation data from participants with real driving experience is crucial.

In this research, we analyzed previously collected data from interviews with Tesla FSD Beta users (Nordhoff & De Winter, 2023) from the meaningful human control perspective. From this data, we gathered information when the participants indicated their perception of trust and safety while describing the behaviour of the automated driving technology. Then, we analyzed that information by classifying whether the behaviour of the vehicles was *tracking* the users' reasons, along with the corresponding perception of trust and safety.

5.3 Related work

5.3.1 Meaningful human control of automated driving systems

Santoni de Sio & van den Hoven (Santoni de Sio & Van den Hoven, 2018) provide a comprehensive philosophical account of two conditions for a system to be under meaningful human control. First, the *tracking* condition requires that a system should be capable of responding to relevant moral, strategic, and intentional reasons of humans in the environment where the system operates. Second, the *tracing* condition implies that a system should be designed in a way that allows tracing back the outcome of its operation to at least one human in the loop of control. The tracking condition has been operationalized for partially automated driving systems (Mecacci & Santoni de Sio, 2020), connecting strategic and tactical reasons with Michon's classical theory of the driving task (Michon, 1985). In this approach, the tracking condition is satisfied when the behaviour of the automated driving system aligns with the moral (e.g. respecting regulations), strategic (e.g., going home), tactical (e.g., overtaking), and operational (e.g., steering) reasons of relevant humans (e.g., the driver). For instance, if a driver has a tactical reason to change lanes smoothly, the system should perform a lane change considered smooth by the driver. Findings from (Mecacci & Santoni de Sio, 2020) have then been used as a basis for design guidelines on human-automation interaction in mixed-traffic (Mecacci et al., 2023). Furthermore, (Calvert & Mecacci, 2020) applied these findings to create a quantitative formulation for vehicle control.

5.3.2 User perception of safety and trust in automated driving systems

Studies consistently demonstrate that drivers' perceptions, particularly their perceptions of safety and trust, significantly influence their willingness to adopt automated driving systems (Ljubi & Groznik, 2023; Montoro et al., 2019). However, such studies have so far been mostly limited to driving simulator experiments and public surveys (Koglbauer et al., 2018; Bellet et al., 2022), which warrants caution in assuming that the results accurately reflect real-world driving experiences. A systematic review of driving simulator experiments revealed variations in the representations (Wynne et al., 2019). Furthermore, many survey studies primarily relied on drivers' imaginative perceptions and expectations of automated vehicles, lacking practical experience and detailed knowledge about the vehicles' functionality (Lee et al., 2023). Survey studies also have disadvantages in terms of information access, reliability, and validity (Burcu,

2000). For this reason, in this work we focus on semi-structured interview data collected previously (Nordhoff & De Winter, 2023) that avoids these issues but also provides in-depth insight into the tactical reasons of users.

5.4 Methodology

5.4.1 Vehicle behaviour

In this work, we connect the concept of meaningful human control to perceived safety and trust via the *behaviour* of automated driving systems. According to (Nordhoff & Hagenzieker, 2024), automation capabilities, including behaviours such as lateral-longitudinal control and collision avoidance, influence the level of perceived safety and trust. These behaviours closely parallel the behaviours of a system that should adhere to the tracking condition, aligning with relevant human reasons. To the best of our knowledge, no existing automated driving systems have been designed with the concept of meaningful human control in mind. Nevertheless, any automated driving system inherently satisfies the tracking condition to some extent (Mecacci & Santoni de Sio, 2020). Therefore, our focus in this paper will be on the tracking condition. When we refer to reasons being tracked or not tracked in subsequent sections, it implies whether the vehicle can fulfill the expected tactical or operational reasons of the driver.

The vehicle behaviour analyzed in this research is related to the driving tasks of vehicles equipped with automated driving technology. When drivers discussed how the vehicle performed the driving task, it indicated whether the vehicle tracked their reasons or not. Given that the subject of this research is SAE Level 2 vehicles, our focus was on examining the driving tasks that automated driving technologies can perform according to the SAE standard (SAE International, 2021). According to this standard, vehicles should have driving assistance features, such as adaptive cruise control and lane centering, at the same time. However, most of the participants did not explicitly mention the names of these features when explaining the behaviour of FSD Beta and standard Autopilot and their perceived safety and trust. Therefore, we only highlighted driving tasks that encompass these features, namely steering, braking, and accelerating.

5.4.2 Data

In this study, we utilized transcribed conversation data from (Nordhoff & De Winter, 2023). Our analysis covered aspects that were not discussed in previous research that used this data (Nordhoff & De Winter, 2023; Nordhoff et al., 2023; Nordhoff & Hagenzieker, 2024). The data comprised responses from 103 respondents in the FSD Beta program, collected through interviews using a semi-structured protocol that included a total of 35 questions, comprising both open-ended and closed-ended questions. The questions consisted of five sections: general experience, perceptions of vehicle operation, perceptions of safety, exploration of trust level, and typical vehicle usage. The interviews were conducted via Zoom and the participants were recruited through social media. On average, each interview lasted 78 minutes and yielded approximately 12,200 words. Following quality checks, which involved the removal of non-English interview data and addressing missing transcriptions, only 99 interview data were deemed suit-

able for further analysis in this research. We only used the answers to the open-ended questions because the purpose was to look for reasons that were only mentioned when the participants explained them. Throughout the interviews, both video and conversation were recorded. The interviewer's role was limited to reduce the possibility of bias.

5.4.3 Data processing

Based on the collection of documents where the transcribed conversation data was saved, we further processed the data so that it could be used for further analysis in our keyword search algorithm. We processed each transcript by applying a series of text preprocessing steps: removing newline characters, adding spaces between digits and alphabets, and combining these operations to create a cleaned version of the conversation. The text was then tokenized using the '*word_tokenize*' function in NLTK (https://www.nltk.org) to break it into individual words. Subsequently, a text cleaning function was applied, removing short words, eliminating spurious characters, transforming letters to lowercase, and removing stopwords and specific words defined in the process, such as the names of participants. The final step involved lemmatization, which normalizes the tokens using NLTK's WordNet lemmatizer. To illustrate, consider a sample text such as "*So changing lanes and avoiding blind spot collisions is very, very good here. I'm maintaining speeds and safe speeds.*". After applying the data processing, the lemmatized tokenized data results will be formatted as a list: [*'changing', 'lane', 'avoiding', 'blind', 'spot', 'collision', 'good', 'maintaining', 'speed', 'safe', 'speed'*].

5.4.4 Seed words and keyword search

To identify instances in the transcribed conversation data where interview participants mentioned the driving tasks and their perceived safety and trust, we employed seed words and a keyword search algorithm with the data. Seed words were defined as individual terms associated with driving tasks, perceived safety, and trust. For driving tasks, we defined the seed words as the verb corresponding to each driving task, except for steering. The choice of the seed word "steer" was often connected to the expression "steering wheel," which was not our focus. Therefore, we sought alternative terms describing the steering process, namely 'change lane' and 'keep lane'. Defining seed words is an iterative step, aligning with Watanabe & Zhou (2022)'s suggestion to derive seeds based on the researcher's subject knowledge. For perceived safety and trust, we adapted the seed words used by Nordhoff & Hagenzieker (2024). Subsequently, our keyword search algorithm, which operates by using the defined seed words as keywords, was employed to identify instances in each conversation where participants mentioned these seed words. We chose keyword search over manual inspection because relevant content could potentially span across various answers to different questions, and the considerable length of the conversations made manual inspection impractical. The seed words and the algorithm can be found in Table 5.1 and Algorithm 5.1. In Table 5.1, the use of 'and' indicates that both words must be present for the keyword search algorithm, while the use of 'or' indicates that the presence of either one is sufficient. As an example, assume we have a list of lemmatized tokenized data: *['vehicle', 'good', 'brake', 'really', 'hard', 'try', 'make', 'left', 'turn', 'reason', 'turn', 'blinker', 'quick', 'definitely', 'make', 'feel', 'unsafe', 'house', 'spot', 'signal', 'make', 'better', 'before']*. We are looking for words discussing perceived safety and trust, particularly

those related to braking. We set *'brake'* as the seed word for braking and *'safe,' 'happy,' 'relax,' 'comfort,' 'glad,' 'trust,' 'distract,', 'rely,'* and *'trustworthy'* as seed words for perceived safety and trust. The algorithm goes through the list, searching for occurrences of these seed words. When it finds one occurrence in one of the seed words, for instance, the word 'brake', it will expand the search over the next 20 words, resulting in a new list of words *['brake', 'really', 'hard', 'try', 'make', 'left', 'turn', 'reason', 'turn', 'blinker', 'quick', 'definitely', 'make', 'feel', 'unsafe', 'house', 'spot', 'signal', 'make']*. Then, the algorithm checks for any words related to the seeds of perceived safety and trust, and it identifies the word 'safe' inside the term 'unsafe'. Now, there are two occurrences of seed words in the list, and the list will be retrieved.

Algorithm 5.1 Keyword Search Algorithm

Require: Seed words $seeds$, tokenized data $tokenList$, buffer size $bufferSize$
Ensure: Retrieved list L

$bufferSize \leftarrow 20$
$L \leftarrow [\,]$ ▷ empty list
$threshold \leftarrow |seeds|$ ▷ number of distinct seed words
for $index \leftarrow 1$ to $|tokenList|$ **do**
  $token \leftarrow tokenList[index]$
  if $token \in seeds$ **then**
    $end \leftarrow \min(index + bufferSize,\ |tokenList|)$
    $tokenBuffer \leftarrow tokenList[index : end]$
    $seedCount \leftarrow \sum_{s \in seeds} \mathbf{1}[s \in tokenBuffer]$
    if $seedCount > threshold$ **then**
      append $tokenBuffer$ **to** L
    end if
  end if
end for

Table 5.1: Seed words

Category	**Sub-category**	**Seed words**
Driving tasks	Accelerating	`accelerate`
	Braking	`brake`
	Lane changing	`lane` and `change`
	Lane keeping	`lane` and `keep`
Perceived safety	–	`safe` or `happy` or `relax` or `comfort` or `glad`
Trust	–	`trust` or `distract` or `rely` or `trustworthy` or `comfort`

After defining seed words and implementing keyword search, we qualitatively classified perceptions by analyzing the word context surrounding the keywords in the transcribed conversations. The classification process began with an initial assessment conducted by the first author. Subsequently, to minimize subjectivity in categorization, discussions were held with the co-authors to ensure alignment with others' perspectives. These discussions involved verifying whether the systems tracked participants' reasons during actions such as braking and determining whether the braking process indicated positive or negative perceptions of safety and trust.

5.5 Results

After performing the keyword search algorithm and qualitative classification, we obtained the results displayed in Table 5.2. The first column represents the driving tasks of the automated driving systems. The second column classifies perceived safety and trust into four categories: safe, unsafe, trust, and lack of trust. The numbers in the third and fourth columns indicate whether the vehicle satisfied the tracking condition of meaningful human control (i.e., tracked or did not track the participant's reasons).

In addition to summary statistics (Table 5.2), we qualitatively analyzed the instances of participants' perceptions of driving tasks performed by automated driving systems, along with details regarding whether their reasons were tracked or not. As the primary focus of this research was to investigate the alignment between tracking and perceived safety and trust, we specifically chose parts of the results that fall within the same category for each perception for further analysis. We developed three categories by examining the pattern of reasons tracked and not tracked for each perception. The first category, 'Controversy,' includes perceptions where both reasons were tracked and not tracked. The second and third categories, 'Reasons mostly tracked' and 'Reasons mostly not tracked,' show perceptions where the majority of reasons were either tracked or not tracked.

Table 5.2: Driving task and user perception analysis results

Driving tasks	**Perception**	**Reasons tracked**	**Reasons not tracked**
Accelerating	Safe	2	2
	Unsafe	0	8
	Trust	1	0
	Lack of trust	0	4
Braking	Safe	19	7
	Unsafe	0	21
	Trust	8	2
	Lack of trust	1	5
Lane changing	Safe	9	4
	Unsafe	0	5
	Trust	4	0
	Lack of trust	0	0
Lane keeping	Safe	11	0
	Unsafe	0	3
	Trust	9	0
	Lack of trust	0	0

5.5.1 System perceived as safe

The results showed that when the participants perceived the driving tasks of accelerating, braking, and lane changing as safe, there were instances where their reasons were tracked and not tracked.

Controversy: lane changing

The participants mentioned thirteen instances where they felt safe with the systems. Nine instances indicated that the systems tracked participants' reasons, while the rest did not. One participant emphasized that significant improvements in automatic lane changes contributed to their sense of safety. Previously, their vehicle would perform several unsafe lane changes, but now it has become as safe as their own driving. Another participant simply did not want to be bothered with lane-changing, as they believed that the systems worked properly and safely.

- *"I do feel safe when it's doing automatic lane changes, and that's a massive improvement since when it first was released, it used like almost every autopilot lane change when it first came out. Would be cutting someone off. Pulling into a lane with someone rapidly approaching or some other form of unsafe lane change now..The automatic lane change feature is.. about as safe as I am. (R007)"*
- *"I don't want to be bothered with changing lanes.. So from my perspective, let the car do it. I know the car is safe, it will do it properly. (R035)"*

Four participants mentioned that they felt safe even though their reasons were not tracked. One of them described how the vehicles kept their turn signals differently from human drivers, but they believed it was because the system was a better driver. Another participant felt safe and shared an occasion where they were impressed with how the system took a lane. They were initially unsure if it was a faster lane, but it turned out to be correct several seconds later.

- *"I feel safe when I am on the freeway.. it's s better driver than I am at taking turns on the freeway.. Changing lanes.. I like how it keeps his turn signal on all the way through the entire lane change and sometimes human drivers will just do a couple blinks. (R043)"*
- *"I do feel safe with autopilot.. but I feel like it's more of a clever thing like it makes really interesting decisions.. like changing into a lane that doesn't yet appear to be faster, but it says it's changing into a faster lane and then 10 seconds later it is the faster lane.. autopilot picked the correct lane and it does it so many times it's like it can't be coincidence. (R074)"*

Controversy: braking

Participants mentioned nineteen instances where the system tracked their reasons. One participant emphasized feeling safe due to the automated driving system's ability to slam the brakes faster than they could react. Another participant indicated a feeling of relaxation, which is linked to the perception of safety, as it reduced repetitive driving tasks during traffic jams. For example:

- *"Completely safe.. You know, it slammed on the brakes when the guy in front of me slammed on the brakes, faster than I could. (R084)"*
- *"It can just kind of help you relax a bit.. one time I was stuck in a traffic jam and autopilot was nice because.. I didn't have to.. put the gas on and then put the brake on and then put the gas on in there.. like over and over and over again. (R059)"*

However, seven instances were perceived as safe situations even when participant's reasons were not tracked. Two participants felt safe in braking scenarios, like sudden brakes or brakes that were too slow in their perception, because they could quickly take over. For example:

- *"I keep my foot almost always on the gas pedal so that if it brakes suddenly, I can quickly override it.. That's how I feel very safe. (R047)"*
- *"So if I'm coming up on a stop sign and it's not slowing down soon enough for what I would expect, then I will manually hit the brakes and disengage.. I've never felt unsafe though and using it. (R087)"*

Reasons mostly tracked: lane keeping

When the participants perceived lane keeping as safe, all instances showed that their reasons were tracked. Participants mentioned eleven instances where the automated driving system's lane keeping tracked their reasons and gave them safe perceptions. One participant described that they felt safe because the systems reduced their main workload so that they did not have to pay attention to lane keeping. Another participant mentioned that the systems did a very good job of lane keeping. For example:

- *"I feel like Autopilot makes me safer because it reduces my workload.. in terms of lane keeping mainly.. It makes me feel more comfortable as a driver just because I don't have to pay attention to the lines on the road. (R024)"*
- *"I feel safe when autopilot and FSD beta..they're not perfect, but yes. They do a very good job of keeping you in your lane. (R099)"*

5.5.2 System perceived as unsafe

When the participants perceived all driving tasks as unsafe, all instances showed that their reasons were not tracked.

Reasons mostly not tracked: lane changing

When the participants explained the situations in which they felt unsafe while changing lanes, the automated driving systems did not track their reasons at all. One of the participants mentioned that they did not feel safe because the system did not have human-like behaviours. They further emphasized that it was too fast or too close to other vehicles at stop signs or traffic lights. Another participant highlighted that they could feel safer if the systems could adapt to how humans drive and incorporate a more human-like feeling into them.

- *"How do you feel when you feel safe or unsafe?.. once they start getting to the point where it has more human like behaviours and doesn't have those.. conflicts in what it's seeing and hitting the brakes or making a turn, you know, changing lanes. (R081)"*

- *"Having a limit on acceleration when changing lanes.. might make sense in kind of a textbook way of safety driving.. it does not match up with the way that humans drive and how you should be adapting to how humans drive.. putting in some more.. natural feeling or more human feeling.. would make me feel safer." (R050)*

Reasons mostly not tracked: accelerating

No reasons were tracked when the participants indicated that they felt unsafe with the automated driving system's acceleration. Out of eight instances where they described feeling unsafe, their reasons were not tracked. One of them mentioned that the vehicle did not feel like a safe driver because it did not blend its speed to accelerate with the flow of traffic. Another participant felt unsafe when merging on the highway because the vehicle did not do a good job of accelerating to match the other vehicle.

- *"It would be safer for the car to accelerate more just with the flow of traffic.. just being able to kind of blend in with the drivers around you, I think as part of being a safe driver. And so since our still situations where the car will not speed up when appropriate or when safe, I'm going to say that there are safety issues. (R050)"*
- *"But when I'm merging on to the highway. That's when I feel the most unsafe because it it will get on to the on ramp. Accelerate to match the speed of the other vehicles, which it doesn't really good job of. (R074)"*

5.5.3 Trust

Our results indicate that participants had trust in the automated driving system's braking in eight instances when their reasons were tracked and in two instances when their reasons were not tracked.

Controversy: braking

The participants described instances when they had high and low levels of trust in the braking experience. When they mentioned that they had trust in the automated driving systems, their reasons were tracked eight times, and two times their reasons were not tracked. Things that made the participants trust the automated driving systems were their capability to brake according to the traffic rules and brake to stop better than themselves.

- *"One advantage of the larger Autopilot is that it can automatically stop at traffic lights. That works pretty well too. It's comfortable. (R047)"*
- *"Stays safe distances I have been in on-air state whenever they just slammed on their brakes and it has stopped very well. Probably better and I would have done. I would have panic stopped and it stopped perfectly. Maintains a good distance. Ohh. So yes, I trust it. (R062)"*

However, there were also instances when they had trust even though the vehicle did not track their reasons. One participant mentioned that even though they intervened in some instances where they felt unsafe, they felt more comfortable the more they drove because the level of perceived safety changed over time. Another participant described a situation where they hit the brake to avoid a long truck but still felt comfortable with the system. They attributed this comfort to the benefit of feeling less tired, even though they still needed to monitor the system.

- *"So whenever I feel unsafe or even uncertain, I'll.. tap the brakes or move the stock up.. it's not difficult for me to do that.. I go in and out of FSD all the time, and that's how I handle this issue of when I don't feel safe.. As your perceived safety changed over time.. the more I drive it, the more comfortable I get." (R103)*
- *"That truck tried to turn into my lane and the car. I didn't realize that it was a really long trailer. So I had to just slow it down, hit the brake, but so you can get pretty comfortable with it and pretty relaxed. I would say you know, so still watching, but it's it actually you're less tired when you get where you're going." (R100)*

Reasons mostly tracked: lane changing

Nevertheless, when the participants had trust in the driving tasks of accelerating, lane changing, and lane keeping, all instances indicated that all of the participants' reasons were tracked. During lane-changing situations, participants mentioned four instances where their reasons were tracked and they had a higher level of trust. One participant initially faced trust issues with the system because it frequently placed them in the wrong lane. However, after an update that required confirmation before lane changes and smooth experiences, the participant's trust started to build. Another participant described a trust-building process with a software upgrade, allowing the system to perform complex manoeuvres in city traffic.

- *"You frequently get in the wrong lane.. And so I did all the driving through there. Uh, because I didn't trust Autopilot.. but by then I gained some confidence in the lane changing.. so I let it tell me when to change lanes and I just confirmed it.. do the lane changing. It was the smoothest trip I've ever taken. (R010)"*
- *"I got the full self-driving upgrade software.. and then I had to get used to it. Changing lanes had to get used to it, doing on ramps and off ramps.. And then I went to FSD beta. And then now I've got to be able to.. trust and get used to the car doing city traffic and these very complex interceptions, these complex intervals (R048)"*

5.5.4 Lack of trust

When participants expressed lower trust in the automated driving system's braking, one instance showed that their reasons were tracked, while five instances indicated that their reasons were not tracked.

Reasons mostly not tracked: braking

Participants who felt lack of trust described that the systems failed to track their reasons. They mentioned that they did not trust the automated driving systems because the systems could make unexpected wrong decisions very fast, such as braking when there was a pedestrian on the crosswalks. Another participant described that they did not trust the systems when entering a highway because the systems did not brake when there was not much of a gap.

- *"I don't fully trust it when it's active.. it can just do something wrong very fast.. you just do not expect it. For example, if it sees a pedestrian.. on the crosswalk, it'll kind of just slam on the brakes. (R065)"*
- *"You're gonna turn onto that highway.. It will creep forward.. it'll sometimes creep into the other highway. So you have to hit the brakes, or it'll appear to start going when there's not much of a gap, so you slam on the brakes again. So I don't trust it. (R010)"*

However, in one instance, the systems tracked their reasons. The only participant who felt lack of trust, even though the systems tracked their reasons, mentioned they admitted Autopilot may react and brake faster than a human, but they felt uncertain because they still needed to rely on themselves.

- *"I think the Autopilot may even be better than a human, and if we take emergency braking, then it can probably react and brake faster than me or certainly close to that, but.. you have to assess the situation, then it's uncertain, and that's why I always have to rely on myself. (R015)"*

5.6 Discussion

5.6.1 Alignment between tracking and perception of safety

Perceived safety

Our results revealed that participants could feel that lane-changing and braking behaviours of the vehicle were safe, even when the reasons were not being tracked (the 'controversy' category). Additionally, we also found the 'reasons mostly tracked' category, where participants perceived lane-keeping tasks as safe, and all their reasons were tracked. Participants who felt safe with lane changing observed better vehicle performance in tracking their reasons, noting that actions like acceleration, deceleration, and lane changes were as safe as manual driving. This aligns with Koglbauer et al. (2018), where positive experiences with vehicle adaptation led to increased perceived safety and trust. The system's performance in lane-changing, braking, and lane-keeping tasks influenced participants' safety perceptions by meeting expectations and providing a relaxed driving experience with less difficulty and workload. According to Xu et al. (2018), reliable experiences with self-driving vehicles increased trust, perceived usefulness, and perceived ease of use. This likely explains why participants believed the system was safe.

On the other hand, feeling safe doesn't always mean that the automated driving systems tracked all their reasons; there were instances of lane-changing and braking tasks where tracking failed. One participant noted that braking was too slow for them, indicating their expected deceleration reason was not tracked. However, they felt safe as long as they were ready to take over. When the driver realized operational reasons for deceleration were not fulfilled, they often took control. The participant also reported no difficulties with the takeover process. Ma & Zhang (2021) found that aggressive drivers were more likely to take control when driving defensively programmed automated vehicles (AVs), potentially explaining the frequent initiation of takeover processes by drivers. The study also revealed that perceived safety levels were similar when aggressive drivers operated both aggressive and defensive AVs. Despite this, the safety score remained higher than when defensive drivers operated aggressive AVs, potentially explaining the continued feeling of safety.

The other instances indicated that the vehicle acted differently from what they expected, but it still led to perceived safety because they believed it was a better driver. The experience of many unexpected manoeuvres that turned out to be correct decisions also influenced their belief. The more the vehicles behaved according to the expected reasons, the more they gained trust from the driver. This trust would make people perceive the system as safe, even if it fails to track their reasons. This is in line with the findings on the effect of reliable experiences with trust that the system is safe (Xu et al., 2018). Thus, we found indications that tracking driver expectations, such as performing safe manoeuvres like lane-keeping, braking, and lane-changing like human drivers, positively correlated with perceived safety. However, failure to track reasons for lane changing and braking was not necessarily associated with perceived lack of safety; it depended on other factors like the numerous reliable experiences, the driver's trust level, and ease of taking over control.

Perceived lack of safety

We observed instances where participants felt unsafe during accelerating and lane-changing experiences. Interestingly, in all these instances, the systems failed to track their reasons. Participants noted that the vehicles did not drive in a manner consistent with human behaviour. According to Peng et al. (2022), human drivers express a higher comfort level with systems that mimic human driving. In this research, we linked the term 'comfort' to the perception of safety and trust, aligning with our observations. Another instance occurred when the vehicle failed to adjust its speed to match the surrounding vehicles during merging or while on the road. It seems that when the vehicle does not track the driver's reasons in situations involving other vehicles that might lead to dangerous situations, it contributes to the perception of being unsafe. This observation aligns with findings from Borowsky et al. (2010), which suggest that drivers tend to intensely pay attention to potential dangers with other vehicles in specific traffic situations, such as merging roads. Our findings suggest that the failure of automated driving systems to track drivers' reasons for human-like acceleration and lane-changing behaviours in potentially dangerous traffic situations can contribute to perceived lack of safety.

5.6.2 Alignment between tracking and level of trust

Trust

Our findings highlighted that participants could exhibit trust in vehicle's behaviour during braking tasks regardless of whether their reasons were tracked (the 'controversy' category). Additionally, we observed another category: 'reasons mostly tracked,' where participants expressed higher trust in lane changing, and all of their reasons were tracked. We further examined situations where participants demonstrated a higher level of trust in both braking and lane changing, analyzing scenarios where their reasons were tracked. In these instances, participants emphasized that their positive experiences with the system's performance in intended situations contributed to a higher level of trust. Again, this is in line with the finding from Koglbauer et al. (2018). Furthermore, the transparency experienced during lane changes, where the driver could anticipate the direction of the vehicle's lane change and had the chance to confirm it, significantly contributed to building the driver's trust. This aligns with a study by Nordhoff & Hagenzieker (2024) that indicates transparency positively impacts trust levels. Additionally, it supports findings from Detjen et al. (2021), suggesting that using displays to indicate manoeuvre intentions increases overall transparency in the driving experience.

In contrast, there were instances where participants expressed trust even though the automated driving systems failed to track their reasons for braking tasks. Despite encountering situations where the vehicle made them feel unsafe, the participants maintained their trust in the system due to the ease of taking over and the relaxed experience it provided. When the driver took over control of the system, we expected they did that to reduce the perceived risk they faced. As they had ease of taking over, they could reduce the perceived risk. According to Zhang et al. (2019) and He et al. (2022), perceived safety risk has a negative correlation with trust. Thus, these findings might justify why the driver kept feeling safe because they could take over control to reduce the perceived risk, thereby increasing their trust level. Furthermore, the relaxed experience with the system suggests that participants have previously had positive experiences, contributing to a higher level of trust (Xu et al., 2018). These situations could recover the trust that might temporarily decline during the takeovers or failure of the systems to track the participant's reasons (Kraus et al., 2020).

We found indications that tracking driver expectations, including improved manoeuvre execution and human-like performance in intended traffic situations in lane-changing and braking tasks, positively relates to trust. Additionally, transparency, which does not relate to tactical or operational reasons, was also positively associated with trust. Notably, the failure to track reasons for braking tasks did not always result in the lack of trust; factors such as positive experiences and the ease of taking over control played a crucial role here.

Lack of trust

Our analysis showed situations where participants reported a lower level of trust in braking tasks; their reasons were always not tracked in such cases. Participants expressed concerns that the system could make unexpected decisions, potentially leading to collisions with other vehicles, thereby reducing their trust. In one contrasting case, a participant had lack of trust, but their reasons were tracked. They believed the system could outperform humans in reaction

time and braking. However, the uncertainty introduced by the need to assess situations led them to consistently rely on themselves. Manufacturers of automated driving systems explicitly instructed FSD Beta program users to maintain constant attention and be ready to act at any time (Nordhoff et al., 2023). This requirement might explain their lack of trust, as it implies reliance on personal judgment.

5.6.3 Limitations

First, the data in this research focused on drivers' subjective perceptions which may not accurately reflect the actual on-road situation (e.g., findings from von Stülpnagel & Lucas (2020) indicate that subjective risk can significantly differ from actual crash risk). Future research should investigate telemetry data in addition to subjective reports. Second, we limited the tactical and operational reasons for relevant human agents only to the driver's perspective. However, many other agents (e.g., vehicle manufacturers and other road users) could and should have meaningful human control over automated driving systems (Mecacci & Santoni de Sio, 2020). We recommend further research to evaluate the tracking of the reasons of other relevant human agents for a more holistic analysis. Third, our results (e.g., Table 5.2) might not have fully captured the entirety of the participants' responses. The use of seed words may restrict instances that are essentially the same but articulated with different words. We recommend incorporating a more extensive set of alternative seed words, especially in the context of driving tasks, for more comprehensive results. Fourth, our findings are limited by the small number of participants in the FSD Beta program. Future research should aim for a more diverse sample encompassing a wider range of manufacturers and participants to better represent the population. Finally, the qualitative nature of our evaluation presents challenges in scalability, particularly when confronted with a larger database of interviews. Given the expansive volume of data, our current methods may prove impractical to implement.

5.7 Conclusions

This study investigated the alignment between tracking component of meaningful human control and user perception of safety and trust. Successfully tracking drivers' reasons for driving tasks, such as safe manoeuvres, performance improvement, and transparency, positively influenced the perception of safety and trust levels. However, the failure to track reasons for lane changing and braking tasks did not necessarily result in negative perception of safety and low trust. Factors such as the ease of taking over control and reliable experiences contributed to drivers feeling safe and maintaining trust. Nevertheless, the failure to track drivers' reasons for expected movements and human-like behaviours in potentially dangerous traffic situations was associated with perceived lack of safety and low trust. Our results can help the developers of automated driving technology to design systems that are under meaningful human control and are perceived as safe and trustworthy.

Chapter 6

Subjective Assessment of Meaningful Human Control

This chapter offers an empirical assessment of Meaningful Human Control (MHC) in partially automated driving systems, with particular attention to drivers' perceptions of safety and trust in their interactions with automation. Building on the conceptual and empirical foundations established in Chapter 5, it extends the analysis from perceived safety and trust to a comprehensive assessment of MHC compliance through both the tracking and tracing conditions. Using a dataset of 103 semi-structured interviews with users of Tesla's Autopilot and Full Self-Driving (FSD) Beta systems, the chapter examines how real-world user experiences align with theoretical MHC principles. The analysis uncovers key influences on perceived meaningful human control, such as system reliability, driver vigilance, and moral awareness of responsibility. By operationalising tracking and tracing into qualitative evaluation criteria, the chapter reveals how drivers interpret their supervisory role and how automation design shapes the distribution of responsibility between human and machine.

The conclusion that Tesla's systems do not fully meet MHC requirements should be understood as a diagnostic finding. The value of the analysis lies not in the fact of non-compliance, which could be anticipated given a system design that assigns all formal responsibility to the driver while keeping decision logic opaque, but in the specific dynamics that empirical analysis reveals: which safety features produce tracking gaps and under what conditions, how trust-induced complacency erodes tracing compliance at the behavioural level despite drivers' formal awareness of their supervisory responsibility, and why assigning formal responsibility alone is insufficient to achieve meaningful human control in practice. This chapter advances the evaluation aspect of the thesis by extending the user-centred analysis of MHC from tracking alone to the combined assessment of tracking and tracing.

This chapter is based on the following paper, published in *Transportation Research Part F: Traffic Psychology and Behaviour*:
Suryana, L. E., Nordhoff, S., Calvert, S. C., Zgonnikov, A., and van Arem, B. (2024). Meaningful Human Control of Partially Automated Driving Systems: Insights from Interviews with Tesla Users. Transportation Research Part F, Vol. 113, pp. 213–236.

6.1 Abstract

Partially automated driving systems are designed to perform specific driving tasks—such as steering, accelerating, and braking—while still requiring human drivers to monitor the environment and intervene when necessary. This shift of driving responsibilities from human drivers to automated systems raises concerns about accountability, particularly in scenarios involving unexpected events. To address these concerns, the concept of meaningful human control (MHC) has been proposed. MHC emphasises the importance of humans retaining oversight and responsibility for decisions made by automated systems. Despite extensive theoretical discussion of MHC in driving automation, there is limited empirical research on how real-world partially automated systems align with MHC principles. This study offers two main contributions: (1) an empirical evaluation of MHC in partially automated driving, based on 103 semi-structured interviews with users of Tesla's Autopilot and Full Self-Driving (FSD) Beta systems; and (2) a methodological framework for assessing MHC through qualitative interview data. We operationalise the previously proposed tracking and tracing conditions of MHC using a set of evaluation criteria to determine whether these systems support meaningful human control in practice. Our findings indicate that several factors influence the degree to which MHC is achieved. Failures in tracking—where drivers' expectations regarding system safety are not adequately met—arise from technological limitations, susceptibility to environmental conditions (e.g., adverse weather or inadequate infrastructure), and discrepancies between technical performance and user satisfaction. Tracing performance—the ability to clearly assign responsibility—is affected by inconsistent adherence to safety protocols, varying levels of driver confidence, and the specific driving mode in use (e.g., Autopilot versus FSD Beta). These findings contribute to ongoing efforts to design partially automated driving systems that more effectively support meaningful human control and promote more appropriate use of automation.

6.2 Introduction

Partially automated driving systems—classified as SAE Level 2 automation—are designed to assist drivers with specific tasks such as steering, accelerating, and braking, while still requiring human drivers to maintain vigilance and be prepared to take control when necessary (SAE International, 2021). The deployment of these systems raises critical questions about the allocation of driver responsibility, especially in unexpected situations. Recent fatal collisions involving Tesla's Autopilot and Ford's BlueCruise have brought these concerns to prominence, particularly in instances where neither the driver nor the system adequately responded to visible obstacles (National Transportation Safety Board, 2017, 2018; Robins-Early, 2024). Currently, manufacturers such as Tesla explicitly assign oversight responsibility to the driver, as clearly stated in official safety documentation, which instructs users to remain attentive and ready to intervene when required (Tesla, 2024a). Consequently, a driver's failure to take timely corrective action may render them liable in the event of a collision.

However, supervising partially automated driving systems presents significant challenges for human drivers (Martinho et al., 2021). Empirical studies involving Tesla Autopilot users have shown that prolonged exposure to reliably performing Level 2 automation often results in "passenger-like viewing behaviours," including extreme cases such as drivers sleeping at the

wheel (Nordhoff et al., 2023). These behaviours illustrate the risks associated with overreliance on automation, leading to reduced attention and increased distraction. This phenomenon echoes Bainbridge's seminal analysis of the "ironies of automation," which demonstrated how automation can undermine operator engagement, foster overdependence, and degrade manual driving skills over time Bainbridge (1983). Supporting this perspective, Banks et al. (2018b) found that drivers responsible for monitoring partially automated systems frequently become complacent, raising concerns about their capacity to intervene effectively. Banks further argued that attributing fault to drivers for failures arising from the design and implementation of Level 2 and Level 3 systems is ethically questionable. Moreover, research indicates that drivers of partially automated vehicles are often held disproportionately accountable for collisions, even in situations where system limitations significantly constrain their ability to respond (Li et al., 2016; Awad et al., 2020; Beckers et al., 2022). These findings underscore ongoing concerns regarding the fair distribution of responsibility in the context of partially automated driving.

6.2.1 Meaningful Human Control

Delegating control to automated systems—those capable of executing tasks with varying degrees of autonomy, ranging from partial to full automation—may give rise to responsibility gaps, in which it becomes unclear which human agent should be held accountable for the outcomes of the system's actions (Matthias, 2004; Santoni de Sio & Mecacci, 2021). To address this challenge, the concept of meaningful human control (MHC) has gained increasing prominence in scholarly debates on responsibility attribution within automated contexts (Santoni de Sio & Van den Hoven, 2018). Originally proposed in relation to autonomous weapon systems (Docherty, 2015), MHC emphasises the principle that humans must retain some degree of control over automated decisions to remain morally and legally accountable for the system's behaviour (Santoni de Sio & Van den Hoven, 2018).

Although MHC was initially formulated in the context of fully automated systems—those functioning without human intervention, such as SAE Level 5 vehicles—it has since been expanded to encompass a broader range of automated technologies, including systems that still require human supervision, such as partially automated driving systems (Mecacci & Santoni de Sio, 2020). This wider applicability is particularly relevant in road transport, where system deployment must consider not only the vehicle's operational capabilities but also the complex nature of the transport ecosystem, including interactions with human drivers, pedestrians, cyclists, infrastructure, and the inherently unpredictable dynamics of traffic and weather.

To enhance understanding of how MHC applies across different levels of vehicle automation, the CCAM Taxonomy provides a useful conceptual framework (Connected Automated Driving, 2024). According to this classification, SAE Level 5 vehicles are fully autonomous and operate without any human intervention. In contrast, SAE Levels 1 to 4 represent varying degrees of automation, each requiring some level of human oversight. For instance, Level 1 systems incorporate minimal automation, such as basic cruise control, while Level 4 systems are highly automated but may still necessitate driver intervention under certain conditions.

Designing automated systems in accordance with the principles of meaningful human control (MHC) is essential for addressing responsibility gaps—particularly in contexts where ethical decision-making depends upon clearly defined parameters for human intervention and ac-

countability (Cavalcante Siebert et al., 2023; Santoni de Sio & Mecacci, 2021). Even when human operators are not directly managing a system's real-time functions, they must retain meaningful control over its behaviour to ensure ongoing oversight and the appropriate assignment of responsibility.

While there is broad consensus in the literature regarding the importance of maintaining MHC in the context of automation (Mecacci & Santoni de Sio, 2020; Cavalcante Siebert et al., 2023; Calvert et al., 2024), there is less agreement on how MHC should be conceptualised and implemented (George et al., 2023; Santoni de Sio & Van den Hoven, 2018; Kwik, 2022; Steen et al., 2023). Despite these differing interpretations, Robbins (2023) identify the framework developed by Santoni de Sio & Van den Hoven (2018) as a valuable foundation for designing systems in line with MHC principles. In their work, Santoni de Sio & Van den Hoven (2018) proposed a philosophical framework through which systems can be evaluated for meaningful human control, outlining two key conditions that must be satisfied: *tracking* and *tracing*.

The *tracking* condition requires that automated systems respond appropriately to the relevant reasons of the human agents involved in their design and deployment. These "reasons" can be understood as expectations—that is, the considerations that justify how an automated system ought to behave to align with human values, objectives, and societal norms (Veluwenkamp, 2022). For clarity, the term expectations" will be used throughout this paper to refer to these reasons. In essence, the tracking condition stipulates that the behaviour of automated systems should reflect the expectations of the relevant human stakeholders.

The *tracing* condition, by contrast, requires that automated systems be designed in a manner that enables their actions to be attributed to at least one human agent involved in their development or operation. Tracing presupposes the existence of an individual who not only understands the system's functionality but also accepts moral responsibility for its behaviour.

Taken together, the tracking and tracing conditions proposed by Santoni de Sio & Van den Hoven (2018) provide a foundational conceptual framework for operationalising meaningful human control in cooperative and automated driving contexts (Calvert & Mecacci, 2020), as well as for the broader design and engineering of automated systems, including automated vehicles (Cavalcante Siebert et al., 2023).

6.2.2 Evaluation of MHC over partially automated driving systems

To ensure that MHC principles are upheld, comprehensive assessments of partially automated driving systems are essential. This involves examining how well these systems comply with MHC principles by evaluating both the tracking and tracing conditions (Mecacci & Santoni de Sio, 2020). In the context of automated driving systems, tracking emphasises that the system should respond to the expectations of its designers and the humans who interact with it. For example, if a driver of a partially automated system expects the system to comply with road regulations, the system should behave in accordance with those regulations to effectively track the driver's expectations.

Tracing, on the other hand, requires that at least one human agent involved in the design or operation of the system understands its capabilities and accepts moral responsibility for its actions. In the context of automated driving systems, this means that drivers must be fully aware of their supervisory role and receive adequate training to supervise and intervene when

necessary (Cabrall et al., 2019).

Several studies evaluating MHC have employed the tracking and tracing framework as a basis for analysis. For instance, Calvert et al. (2020b) used the framework to evaluate partially automated driving systems, while Calvert et al. (2021) applied these criteria to assess cooperative vehicles and truck platooning systems.

While these contributions offer valuable insights into the assessment of partially automated driving systems, they primarily rely on hypothetical scenarios or post-incident analyses. Notably absent from much of the existing literature are the subjective experiences of real-world users of such systems. Yet these experiential insights are critical for understanding how users interact with automated driving technologies in everyday contexts. This perspective is essential for ensuring appropriate system use, a core element of both MHC and broader traffic safety considerations (Cavalcante Siebert et al., 2023).

Recent work by Suryana et al. (2024) has begun to address this gap by examining drivers' perceptions of safety and trust in relation to the tracking dimension of MHC. However, comprehensive evaluations of MHC compliance—encompassing both the tracking and tracing conditions—based on users' subjective experiences remain limited in the current literature.

6.2.3 Research Gaps and Objectives

1. **Theoretical Gap**: There is a lack of clarity regarding the application of tracking and tracing methodologies to assess MHC in real-world driving contexts. This issue is particularly critical, as previous studies have demonstrated that drivers frequently exhibit unsafe behaviours—such as complacency, falling asleep behind the wheel, or engaging in non-driving activities—while using automated systems (Wörle & Metz, 2023; Nordhoff et al., 2023). Such behaviours challenge adherence to MHC principles and raise concerns about whether these systems are genuinely under meaningful human control in everyday driving scenarios.

2. **Practical Gap**: Existing assessments of MHC have largely neglected the subjective experiences of drivers operating partially automated systems in real-world settings. For example, the ways in which drivers perceive their supervisory role, interpret system behaviour, and how their perceptions of accountability evolve over time remain insufficiently explored. These experiential factors are essential for determining whether partially automated systems are truly under meaningful human control.

3. **Methodological Gap**: Current approaches to evaluating MHC often overlook critical elements of human-automation interaction. They fail to investigate whether the system's performance consistently aligns with human expectations, or whether drivers fully comprehend their responsibilities and are capable of reclaiming control when necessary. These limitations hinder the effective evaluation of meaningful human control in real-world driving contexts.

To address these gaps, this study applies the framework of MHC to real-world driving contexts, drawing on previously collected interview data from users of Tesla Autopilot and Full Self-Driving Beta systems (Nordhoff et al., 2023). By moving beyond hypothetical scenarios

and post-accident analyses, this research offers a dynamic assessment of MHC in everyday driving situations. It further investigates how drivers perceive their responsibility in supervising automation, the evolution of their trust and safety perceptions, and how they interpret system behaviour—dimensions that have been largely neglected in prior evaluations. Finally, by employing a qualitative methodology that captures the nuanced and context-dependent nature of human-automation interaction, this study provides a more comprehensive approach to evaluating MHC compliance. Collectively, these contributions deepen the understanding of meaningful human control in partially automated driving systems, offering valuable insights for both theoretical development and practical design improvements aimed at enhancing the safety and accountability of driving automation.

6.3 Method

6.3.1 Dataset

This study draws on a dataset comprising 103 semi-structured interviews with active users of Tesla's Autopilot and Full Self-Driving (FSD) Beta systems. The interviews focused on participants' real-world experiences and interactions with these technologies, capturing a broad range of topics including perceived safety, trust, control, and responsibility.

Although participants were not explicitly introduced to the concept of Meaningful Human Control (MHC), the interviews contained numerous responses that align with its theoretical components—specifically, aspects related to tracking (e.g., alignment between system behaviour and human expectations) and tracing (e.g., attributions of responsibility and control). This made the dataset well-suited for retrospective analysis through the lens of the MHC framework.

Details regarding recruitment and study procedures are provided in the following subsections.

Recruitment

The dataset utilised in this study was collected through a recruitment process and interview procedure approved by the Human Research Ethics Committee of Delft University of Technology (ID: 2316). Participants were initially identified through special interest groups related to Tesla vehicles on various social media platforms, including Discord, Facebook, Twitter, Reddit, YouTube, Instagram, Tesla Motors Club, and the Tesla Motors Forum. Snowball sampling was subsequently employed, with participants referring others via email. As Full Self-Driving (FSD) Beta was available only to residents of North America and Canada during the study period, recruitment efforts were predominantly focused on these regions. Eligibility for participation was determined based on self-reported access to Autopilot and FSD Beta. FSD Beta users were individuals selected by Tesla according to safety scores and ownership status. Prior to granting access, Tesla provided the following usage guidelines to FSD Beta users:

"Full Self-Driving is in limited early access Beta and must be used with additional caution. It may do the wrong thing at the worst time, so you must always keep your hands on the wheel

and pay extra attention to the road. Do not become complacent. When Full Self-Driving Beta is enabled, your vehicle will make lane changes off highway, select forks to follow your navigation route, navigate around other vehicles and objects, and make left and right turns. Use Full Self-Driving Beta only if you will pay constant attention to the road, and be prepared to act immediately, especially around blind corners, crossing intersections, and in narrow driving situations. Every driver is responsible for remaining alert and active when using Autopilot and must be prepared to take action at any time. As part of receiving FSD Beta, your vehicle will collect and share VIN-associated vehicle driving data with Tesla to confirm your continued eligibility for FSD Beta feature. If you wish to be removed from the limited early access FSD Beta please email xxx."

Procedure

Interviews were conducted remotely via Zoom, with both audio and video recordings. To ensure consistency and minimise interview bias, a predefined interview protocol was developed using Qualtrics (https://www.qualtrics.com). The link to the protocol was shared with participants via Zoom's chat function at the commencement of the interview, enabling them to follow the questions and progress through them independently. This approach was specifically designed to reduce the interviewer's potential influence on participants' responses.

At the outset of the interviews, participants provided their informed consent to take part in the study. The first section of the interview primarily comprised open-ended questions, focusing on participants' perceptions and experiences with Autopilot and Full Self-Driving (FSD) Beta, including aspects such as feelings of safety, trust, and typical usage (see B). For example, participants were asked to describe situations in which they felt unsafe using these systems, as well as how their trust and safety perceptions evolved over time. The second section of the interview comprised closed-ended questions concerning participants' socio-demographic profile, travel behaviour (e.g., age, gender, education level, frequency of Autopilot/FSD Beta use), and their general attitudes towards traffic safety.

The interviewer's role was primarily observational, intended to minimise bias by allowing participants to navigate the questionnaire independently. However, follow-up questions were posed to clarify responses or explore new themes that emerged during the interview. Participants were also encouraged to skip any questions that had already been addressed. The interviews lasted an average of 78 minutes, resulting in approximately 12,200 words of transcribed data.

To ensure the integrity of the data, four interviews conducted in German were excluded from the analysis to avoid potential issues associated with missing transcriptions or mistranslations that could arise from translating the responses into English. Consequently, 99 of the original 103 interviews were considered suitable for further analysis.

6.3.2 Data analysis

An evaluation framework for MHC was developed to assess whether Tesla's Full Self-Driving (FSD) Beta and Autopilot systems align with the expectations of relevant human agents (tracking), and to what extent individuals involved in the operation and design of these systems understand their capabilities and recognise their moral accountability for the systems' actions

(tracing). This evaluation follows a structured five-step process, as detailed below.

MHC Component Identification

Tracking Component: (1) Human Agents and (2) Their Expectations In this step, we identified the human agents and their safety expectations in order to evaluate tracking alignment. Using the MHC taxonomy as a framework (Figure 6.1), we defined two categories of human agents based on their relationship with the system: *drivers*, classified as proximal internal agents (those who interact directly with the system), and *manufacturers*, classified as distal internal agents (those responsible for designing and regulating system functionality). We also defined their *safety expectations* as tactical expectations, reflecting real-world interactions. Specifically, drivers expected the system to prevent accidents (e.g., by providing collision warnings or automatically applying the brakes in emergency situations), while manufacturers expected the system to comply with safety standards (e.g., meeting regulatory requirements for collision avoidance). This categorisation provided a structured framework for evaluating whether system behaviour aligns with the safety expectations of these human agents.

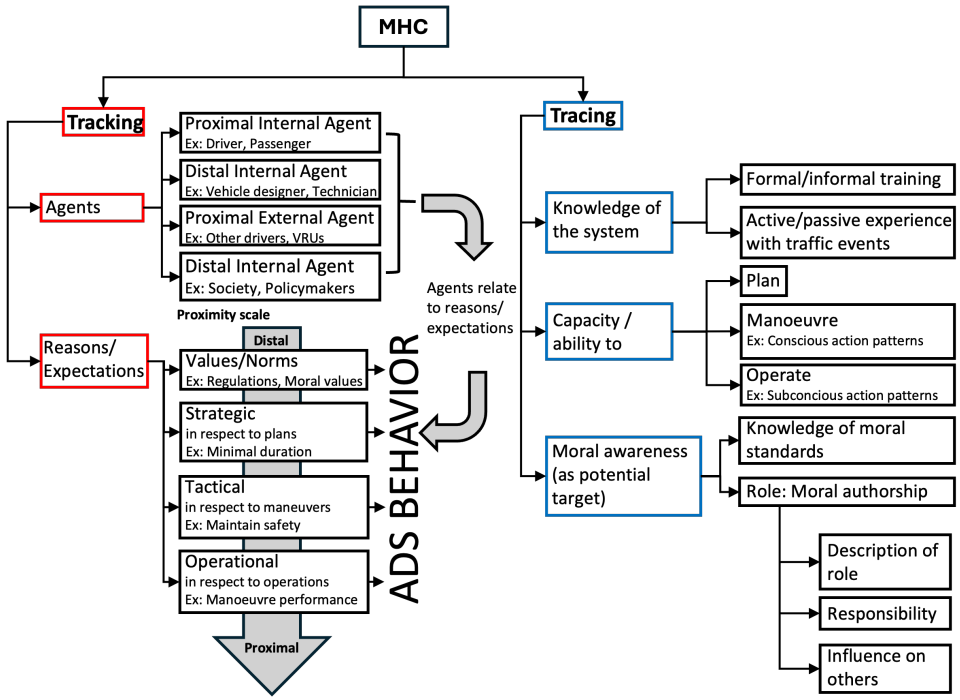


Figure 6.1: Taxonomy of tracking and tracing, adapted from the work of Calvert & Mecacci (2020)

Tracking Component: (3) Features That Influence Vehicle Behaviour In addition to defining human agents and their expectations, this step also identifies the active safety features in Tesla's Autopilot and FSD Beta systems that directly influence the vehicle's behaviour and contribute to meeting safety expectations. These features act as key indicators of how effectively the system tracks and responds to the expectations of human agents.

- *Automatic Emergency Braking (AEB):* Detects vehicles or obstacles in the vehicle's path

and applies the brakes if necessary.

- *Forward/Side Collision Warning (F/SCW):* Alerts the driver to potential collisions with slower-moving or stationary vehicles or obstacles alongside the vehicle.
- *Blind Spot Monitoring (BSM):* Warns the driver when a vehicle or obstacle is detected in the blind spot during lane changes.
- *Lane Departure Avoidance (LDA):* Applies corrective steering to assist in keeping the vehicle within its intended lane.

These features were selected because they directly influence the vehicle's behaviour and are critical for ensuring safety in real-world driving scenarios. By focusing on these features, we were able to assess how effectively the system tracks and responds to the expectations of human agents, thus providing a robust foundation for evaluating alignment with the MHC principle.

Tracing Component To evaluate tracing, it is necessary to identify a human agent who understands the system's capabilities and recognises their moral accountability in the design and operation of the system. In the case of Tesla, we selected the *driver* as the accountable human, as the company explicitly assigns this responsibility to drivers through its operational guidelines (Tesla, 2024b). Prior to engaging Autopilot, drivers must agree to 'keep their hands on the wheel at all times' and to 'remain in control of and responsible for their vehicle at all times.' This requirement highlights the driver's role as the primary human responsible for overseeing system performance and intervening when necessary.

Defining MHC Evaluation Criteria

Tracking Evaluation Criteria To assess whether Tesla's Autopilot and Full Self-Driving (FSD) Beta systems align with human agents' safety expectations, we adapted two evaluation criteria: *Safety of the Intended Functionality (SOTIF)* (International Organization for Standardization, 2019) and *Perceived Safety and Trust (PST)*. The SOTIF framework was selected because evaluating the safety of automated driving systems necessitates a standardised approach, while PST was included because even technically safe systems may fail to align with human expectations if their behaviour is perceived as unpredictable or unreliable.

These criteria were chosen to evaluate both the technical performance of the system and the subjective experiences of drivers. For SOTIF, we employed an adjusted version, termed ad-SOTIF, to compare drivers' descriptions of system behaviour with Tesla's official specifications. If the system's behaviour aligned with the manufacturer's descriptions, it was classified as ad-SOTIF (+); deviations were classified as ad-SOTIF (-).

For PST, we assessed drivers' perceptions of safety and trust based on their interview responses. As our study evaluates safety expectations through driver perceptions, PST serves as a proxy for assessing tactical expectations, as depicted in the tracking taxonomy in Figure 6.1. Trust was incorporated as a criterion due to its strong positive relationship with perceived safety, given that trust is often modelled as a function of perceived safety (Nordhoff et al., 2021). This approach enabled us to capture additional facets of drivers' safety experiences that may not be

explicitly expressed through the word "safe" in interviews, thereby providing a more comprehensive understanding of their perceptions. Positive perceptions, such as feelings of reliability or confidence, were classified as PST (+), while negative perceptions, such as distrust or feelings of risk, were classified as PST (-).

This dual approach enabled us to assess both the technical alignment of the system with its intended functionality and the subjective experiences of drivers, thereby ensuring a comprehensive evaluation of whether the system meets the safety expectations of both drivers and manufacturers.

Tracing Evaluation Criteria To evaluate driver compliance with tracing requirements, we derived three criteria from Tesla's usage guidelines (Nordhoff et al., 2023), as outlined in Section 6.3.1, and aligned them with the MHC tracing taxonomy (Calvert & Mecacci, 2020). These criteria include: *knowledge* (staying alert and keeping hands on the steering wheel), *capability* (performing corrective actions), and *moral awareness* (maintaining operational responsibility).

Table 6.1: Tracing evaluation criteria

Criteria	**Details**
Knowledge	(1) To stay alert and (2) To keep both hands on the steering wheel
Capability	To be able to perform corrective action
Moral awareness	To maintain operational responsibility

These instructions provide the foundation for operationalising the criteria. For instance, the requirement to "keep hands on the wheel" was categorised under knowledge, while "be prepared to act immediately" was mapped to capability. By grounding the criteria in both Tesla's instructions and the MHC framework, this step ensures a structured evaluation of driver compliance with tracing requirements. The criteria are summarised in Table 6.1.

Locating MHC-Related Content

The process of locating MHC-related content in the interview data depends on whether the questions are already aligned with the tracking and tracing evaluation criteria. In instances where the questions were not directly related, additional steps were required to identify and extract relevant content. For example, in our study, the interview questions were not explicitly designed to address tracking criteria, necessitating a more detailed preprocessing and keyword search approach. This method was essential due to the unstructured nature of the data, and keyword searches enabled us to systematically identify segments of the interviews that discussed specific safety features (e.g., Automatic Emergency Braking or Lane Departure Avoidance). In contrast, the tracing criteria were addressed through specific questions in the interview protocol, facilitating the direct extraction of relevant responses. Below, we outline the distinct methodologies used for locating content related to tracking and tracing.

Locating Tracking-Related Content: Data Preprocessing To systematically identify tracking-related content in the interview data, we began by preprocessing the transcribed text, transforming it into word tokens and applying several cleaning steps. Preprocessing ensures that only meaningful content is retained, eliminating noise that could affect the accuracy of subsequent keyword searches. In line with best practices in text analysis (Banks et al., 2018a; Hickman et al., 2022), we used the NLTK Python package (NLTK Project, 2024) to perform the following steps: (1) removal of newline characters and extra spaces, (2) tokenisation into individual words, (3) elimination of short words (< 2 characters) and long words (> 30 characters), numbers, and punctuation, (4) conversion to lowercase and filtering of English stopwords, and (5) lemmatisation to normalise words to their root forms. For example, the sentence *One advantage of the larger Autopilot was that it could automatically stop at traffic lights.* was transformed into the list *['one', 'advantage', 'larger', 'autopilot', 'automatically', 'stop', 'traffic', 'lights']*. This preprocessed dataset was then utilised in subsequent keyword searches.

Locating Tracking-Related Content: Identifying Seed Words To identify tracking-related content, we selected seed words associated with the four active safety features (AEB, F/SCW, BSM, and LDA), which act as indicators of relevant discussions within the interview data. These safety features represent broader themes, while the seed words serve as specific indicators to help locate pertinent content. Adopting a knowledge-based approach (Watanabe & Zhou, 2022), we chose initial seed words—such as "emergency," "braking," "collision," "warning," "blind spot," and "lane departure"—based on their strong association with these safety features.

Once the initial seed words were selected, we applied the pre-trained Global Vector (GloVe) model in Python to enhance the seed word set. The GloVe model, a machine learning technique for generating word embeddings, was employed to identify synonyms and semantically related terms that may not have been initially considered. Details of the pre-trained model and setup instructions are available at the official GloVe project page (https://nlp.stanford.edu/projects/glove/). This enrichment process strengthened the robustness of the keyword search by ensuring that a broader range of relevant terms could be identified within the interview data. For example, the seed word "braking" was enriched with terms such as "deceleration," "traction," and "acceleration," while "collision" was expanded to include "accident," "collide," and "crash." The final enriched seed word set was then used in a systematic search to locate content pertinent to the active safety features. Table 6.2 presents a detailed overview of the initial and enriched seed words, illustrating the outcomes of this process. By combining both expert knowledge and machine learning techniques, this step ensured that the keyword search algorithm effectively identified relevant interview content.

Table 6.2: Seeds related to active safety features

Category	Sub-category	Knowledge-based seeds	Seeds for keyword search
Active safety features	AEB	emergency AND braking	(emergency OR urgent OR disaster OR immediate OR assistance) AND (braking OR deceleration OR steering OR traction OR acceleration)
	F/SCW	collision AND warning	(collision OR accident OR collide OR crash OR head-on OR mishap) AND (warning OR warn OR alert OR indication OR danger OR caution)
	BSM	blind AND spot AND monitor	(blind OR mistaken OR sight OR impossible) AND (spot OR place OR there OR where) AND (monitor OR tracking OR surveillance OR alerting OR evaluation OR utilization)
	LDA	lane AND keeping	(lane OR road OR freeway OR crossing OR roadway OR highway OR ramp) AND (keeping OR kept OR keeps OR putting OR bringing OR maintain)

Locating Tracking-Related Content: Keyword Search Algorithm To systematically identify tracking-related content in the interview data, we applied a keyword search algorithm proposed by Suryana et al. (2024), which utilises enriched seed words to detect relevant segments. This algorithm was applied to the lemmatised tokens generated during the data preprocessing phase. It employed the enriched seed words, generated by the GloVe model, to scan the tokenised data and identify segments where the seed words appeared (see Algorithm 6.1).

The algorithm incorporated logical operators to refine the search process. The 'OR' operator allowed the inclusion of synonyms for the seed words, while the 'AND' operator ensured that paired seed words, as defined in Table 6.2, appeared together within a 20-token sliding window in the lemmatised, tokenised data. The choice of a 20-token window was informed by prior work (Suryana et al., 2024), which demonstrated that this window size effectively captures meaningful contextual relationships between related terms in similar textual analyses. For instance, when applying the knowledge-based seed words for AEB, "emergency" and "braking," the algorithm would detect the occurrence of the word "emergency" in the token sequence and then scan the subsequent 20 tokens to check for the presence of the paired seed word "braking." If both seed words were found within this 20-token window, the corresponding segment of the original transcribed interview would be extracted for further analysis using classifications such as ad-SOTIF(+), ad-SOTIF(-), PST(+), or PST(-).

Locating Tracing-Related Content: Direct Extraction from Interview Responses To identify and extract interview segments related to the tracing evaluation criteria defined in Step 2, we focused on responses to specific questions in the interview protocol: Q25, Q26, Q34, and a question concerning the maintenance of control and responsibility (see Section B). For example, Q25 asked, "Do you typically keep your hands on the steering wheel at all times?" and Q26 asked, "Are you typically fully attentive and alert at all times?", both of which directly relate to drivers' knowledge. Similarly, Q34 ("Do you typically stay prepared to take corrective actions at all times?") provided insights into drivers' capability. The question regarding the

Algorithm 6.1 Keyword Search Algorithm (Suryana et al., 2024)

Require: Seed words (seeds), tokenised data (tokenList), Buffer size (bufferSize)
Ensure: Retrieved list (list)

```
bufferSize ← 20
list ← []
threshold ← ∑(seed ∈ seeds)
for all (token, index) ∈ tokenList do
    if token ∈ seeds then
        tokenBuffer ← tokenList[index : index + bufferSize]
        seedCount ← ∑_{seed∈seeds} (seed ∈ tokenBuffer)
        if seedCount > threshold then
            list ← tokenBuffer
        end if
    end if
end for
```

maintenance of control and responsibility was not explicitly stated in Appendix B, but it could be inferred from drivers' responses. For instance, when drivers read the question aloud and responded with statements such as, "Do not maintain control and responsibility for my car? I strongly disagree," it addressed the moral awareness criterion, ensuring that drivers recognised their accountability for the system's behaviour. The extracted responses were then prepared for qualitative assessment, facilitating a focused and efficient evaluation of drivers' understanding of their responsibilities.

MHC Evaluation

Tracking Evaluation: Content Analysis Following the extraction of tracking-related conversation segments in Step 3, a content analysis was conducted to classify the content based on the evaluation criteria defined in Step 3: ad-SOTIF and PST. The objective of this step was to determine whether Tesla's Full Self-Driving (FSD) Beta and Autopilot systems comply with the tracking requirements of the MHC framework.

For ad-SOTIF, we compared drivers' descriptions of active safety features in the interview data with the intended functionalities outlined on Tesla's official website (Tesla, 2024a). If the descriptions aligned with the manufacturer's specifications, the features were classified as ad-SOTIF (+). For instance, if a driver described Automatic Emergency Braking (AEB) as functioning consistently with Tesla's description (e.g., braking automatically when an obstacle is detected), this was categorised as ad-SOTIF (+). Conversely, if drivers reported discrepancies or failures in system behaviour (e.g., AEB not activating when required), the features were classified as ad-SOTIF (-).

For PST, we evaluated drivers' perceptions of safety and trust based on their interview responses. To assess trust, we identified terms such as "depend," "rely," and "trust," which indicated whether drivers had a positive level of trust in the system. Similarly, terms related to safety, such as "relax," "risk," and "safe," were used to gauge drivers' perceived safety. These terms were selected based on established questionnaires for evaluating trust (Choi & Ji, 2015)

and perceived safety (Xu et al., 2018). If drivers expressed confidence in the system's reliability and felt safe using it, the content was classified as PST (+). For example, a driver stating, "I feel relaxed using Autopilot because it handles most situations well," would be categorised as PST (+). In contrast, if drivers expressed distrust or felt unsafe (e.g., "I don't trust the system to handle sudden stops"), the content was classified as PST (-).

This qualitative analysis enabled us to classify the extracted content into four categories: ad-SOTIF(+), ad-SOTIF(-), PST(+), and PST(-). By combining these classifications, we were able to assess not only the technical alignment of the system with its intended functionality but also the subjective experiences of drivers. This dual approach ensured a comprehensive evaluation of whether the system meets the safety expectations of both drivers and manufacturers, as outlined by the MHC framework. The results of this analysis provided a structured basis for understanding how well Tesla's systems track and respond to human agents' needs, highlighting both areas of alignment and potential gaps.

Tracing Evaluation: Thematic Analysis To evaluate whether drivers' experiences with Tesla's Autopilot and FSD Beta systems comply with the tracing evaluation criteria, we conducted a thematic analysis of their responses. This involved categorising responses into subcategories that reflected drivers' understanding of their responsibilities, knowledge, and capabilities. Following inductive coding principles (Nordhoff, 2024), we performed open coding, reviewing the extracted responses line-by-line to identify recurring themes, such as "keeping hands on the wheel," "monitoring the road," or "feeling responsible for interventions." These themes were then grouped into broader subcategories based on their similarities and distinctions. For instance, responses mentioning "hands on the wheel" and "staying alert" were grouped under a subcategory such as Compliance with Hands-on Requirements. To ensure robustness, we retained only those subcategories mentioned by at least five drivers, as this frequency threshold helped validate the relevance and significance of each subcategory. In cases where a single quote applied to multiple subcategories, each relevant subcategory was assigned a frequency count of one. This systematic approach ensured that the subcategories were both data-driven and representative of drivers' experiences, providing a structured foundation for further analysis.

Illustrative Quotes

Tracking Quotes Representative quotations were selected from the classified content to illustrate the findings. These quotations exemplify each of the four classifications: ad-SOTIF(+), ad-SOTIF(-), PST(+), and PST(-). For each category, excerpts from the interview data were chosen to clearly represent either alignment with or deviation from manufacturer specifications (ad-SOTIF), or to reflect positive or negative driver perceptions regarding safety and trust (PST).

Tracing Quotes To provide concrete examples of the subcategories identified in Step 4, we selected up to three representative quotations per subcategory. Priority was given to quotes that clearly exemplified the theme and reflected common driver experiences.

6.4 Results

This section presents the results of our evaluation of Tesla's Full Self-Driving (FSD) Beta and Autopilot systems in relation to the concept of meaningful human control (MHC), with a specific focus on the tracking and tracing requirements. To illustrate how drivers' feedback aligns with these requirements, we include quotations that reflect their experiences. Each quote is accompanied by the participant ID number for reference. To highlight key insights, we have selected several representative quotes. The results, supported by these quotations, are further discussed in Sections 6.4.1 and 6.4.2.

6.4.1 Tracking Evaluation Results

Our tracking evaluation revealed a varied distribution of safety features across the tracking evaluation criteria (Figure 6.2). For example, in the ad-SOTIF(+) PST(+) category, Lane Departure Avoidance (LDA) and Blind Spot Monitoring (BSM) exhibit a higher percentage distribution compared to Automatic Emergency Braking (AEB) and Forward/Side Collision Warning (F/SCW). Percentage distributions above 80% indicate that LDA and BSM more frequently meet both driver and manufacturer safety expectations, compared to instances where they fail to align with one or both expectations. In contrast, there were no instances of ad-SOTIF(-) PST(+), suggesting that when the features did not perform as intended, drivers never held a positive perception of them.

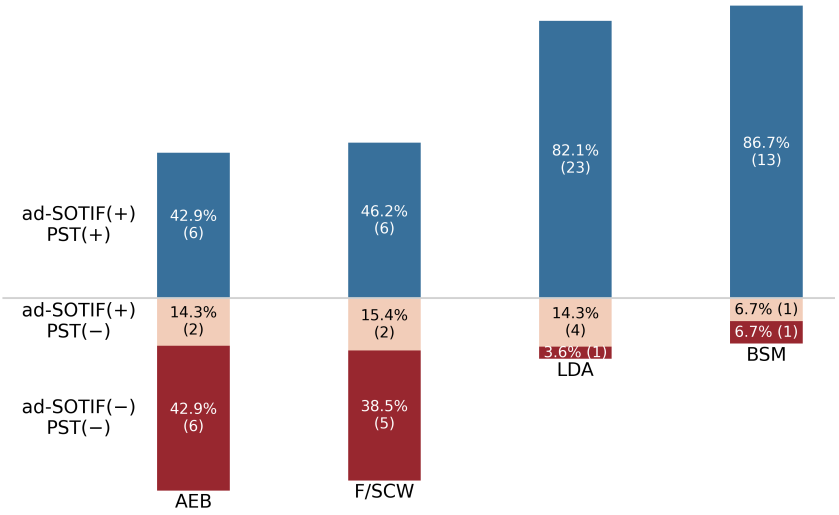


Figure 6.2: Tracking evaluation results for the four active safety features of Tesla vehicles—Automatic Emergency Braking (AEB), Forward/Side Collision Warning (F/SCW), Lane Departure Avoidance (LDA), and Blind Spot Monitoring (BSM)—are presented. For each feature, the stacked bars represent the percentage of instances in which each performance category—alignment with intended functionality (ad-SOTIF) or perceived safety and trust (PST)—was mentioned in the interviews. The numbers in parentheses below the percentages indicate the total number of mentions for each category.

To provide a more detailed insight into the tracking evaluation of the safety features, we classified them into three categories based on the co-occurrence of positive and negative instances of ad-SOTIF and PST. It is important to note that each safety feature could be assigned to multiple categories depending on the variation in user experiences:

- *Inconsistent tracking:* Safety features that were described both as (1) functioning as intended and generating positive perceptions of safety and trust, and (2) not functioning as intended and generating negative perceptions. A feature was assigned to this category when the number of instances with ad-SOTIF(+) PST(+) was comparable to those with ad-SOTIF(-) PST(-), indicating inconsistency in performance and perception.
- *Gap between performance and perceived safety/trust:* Safety features that technically functioned as intended but failed to generate positive perceptions of safety and trust. In such cases, although the features aligned with manufacturers' safety expectations, they failed to meet drivers' expectations. Features with a notable proportion of ad-SOTIF(+) PST(-) instances were assigned to this category.
- *Consistent tracking:* Safety features that not only functioned as intended but also consistently elicited positive perceptions of safety and trust among drivers. Features were included in this category when there was a high occurrence of ad-SOTIF(+) PST(+) and a low occurrence of ad-SOTIF(-) PST(-), indicating strong alignment with both driver and manufacturer safety expectations.

It is important to emphasise that although only a limited number of illustrative quotations are presented in the following sections, each theme was derived from multiple participant responses. The frequency with which each theme was mentioned varied; some were discussed by a larger number of participants, while others emerged less frequently. Moreover, individual responses often encompassed multiple themes, as participants' experiences with the system frequently addressed several aspects of its performance simultaneously. These tracking categories were developed to capture the full range of relevant patterns observed in the data, ensuring that both commonly and less frequently reported experiences were taken into account.

Inconsistent Tracking

Automatic Emergency Braking (AEB) and Forward/Side Collision Warning (F/SCW) were found to align with both drivers' and vehicle manufacturers' safety expectations in certain scenarios, while failing to do so in others. To better understand how AEB and F/SCW can both meet and fall short of these expectations, we analysed participant feedback regarding the performance of these systems. The data revealed several recurring themes in drivers' perceptions of these features:

- **Effective functionality - ad-SOTIF(+) PST(+)**
  Participants mentioned that AEB performs well in detecting vehicles ahead that the driver may not see, thereby preventing potential collisions. One user expressed appreciation for this feature:

"I would have actually hit someone, but they stopped suddenly for some reason, maybe someone was crossing.. I didn't see it, but the system detected it and prevented a collision by performing an emergency brake. It worked really well, and I'm very grateful (R047 - AEB)"

Similarly, the F/SCW feature proved effective in alerting drivers to potential collisions from the front and sides. One participant expressed their appreciation for this feature:

"An autopilot averted a potential accident.. I was very impressed. While driving, my car started.. telling me to take control immediately. I looked in my blind spot, and a car in the next lane veered into mine. I didn't see it, but Autopilot did and reacted right away (R091 - F/SCW)"

- **False positive and false negative errors - ad-SOTIF(-) PST(-)**
  Despite the overall positive performance, drivers also reported instances where AEB and F/SCW did not function as intended. For example, there were cases in which the automated systems failed to respond to debris on the highway and missed alerting the drivers. Both situations are considered false negatives, which led to feelings of unsafety among participants.

  "I think what's unsafe is just right now it only has a front collision warning.. It doesn't gonna detect anything..comes to you from the side.. If it does it.. only if you drive at the really slow speed. (R058 - F/SCW)"

  "If I didn't take over, it would drove right over the piece of wood and probably created a lot of damage that might have caused an accident because hitting at highway speeds, a piece of debris.. Tesla uses cameras as their technology, but you could probably detect better debris and just alert.. like they have some alerts when you're driving if it's uncertain. So they could do that to make it safer (R073 - AEB and F/SCW)"

  Additionally, the system sometimes engaged in "phantom braking," a false positive case in which the brakes were applied without the presence of an actual obstacle. This led to annoyance among drivers:

  Autopilot take care of 99% of driving.. The only issues.. it's not a perfect system.. there are a lot of false positives, particularly in one lane roads where in cars are coming at you fast. It sometimes thinks it's going into your lane and does a phantom brake. In the case.. it.. annoys you by saying, "hey, there's a forward collision warning" when it's not. (R078 - AEB and F/SCW)

- **Software issues - ad-SOTIF(-) PST(-)**
  Drivers also reported unsettling software issues, including automatic warnings upon vehicle reboot and inconsistencies in alarm triggering. These problems contributed to undesirable experiences among drivers:

  "It was.. scary enough that.. a non informed user would not know what to do. Autopilot would constantly disengage my visualization.. rebooting about every three seconds, and

every time it rebooted, a forward collision warning would occur. It would not slow down my car, but it would make like the super loud multiple beeps like I'm gonna hit something. (R055 - F/SCW)"

Overall, our analysis indicates that the AEB and F/SCW systems generally align with both driver and vehicle manufacturer safety expectations under typical driving conditions. Specifically, AEB was frequently noted for its effectiveness in detecting vehicles ahead, while F/SCW was recognised for its ability to alert drivers to potential frontal and lateral collisions.

However, participants also reported instances where these systems failed to meet expectations. These failures included both false positives and false negatives. A commonly reported false positive was phantom braking—where the vehicle applied the brakes without a discernible obstacle. False negatives included failures to detect road debris, side collisions, or to provide timely warnings to the driver. In addition, several participants reported software inconsistencies, such as unexpected system reboots, which further undermined the reliability of the tracking function. These issues suggest that while AEB and F/SCW often perform as intended, limitations remain that affect their consistency and overall effectiveness.

Gap between Performance and Perceived Safety/Trust

Three safety features—AEB, F/SCW, and LDA—were classified under this category[1]. This classification was based on driver responses indicating that these features generally functioned as intended but were nonetheless associated with negative perceptions of safety and trust (PST). Based on participants' feedback, we identified several potential causes for this perception gap, which are outlined below.

- **Premature collision warnings - ad-SOTIF(+) PST(-)**
  One participant described a situation in which their vehicle issued collision warnings for vehicles that were still a considerable distance away and then abruptly applied the brakes. Although this suggests that the warning system was effective at alerting the driver to potential future collisions, the participant felt that these warnings were unnecessary, as the vehicle in front still had sufficient time to complete the turn before the partially automated driving system reached it. This issue led to frustration with the system:

  "One of the annoying things.. this is a little tiny bit as safe.. You're from Holland.. So you are on the right side of the road.. So when you're driving.. someones turning left in front of you and they're like way ahead.. like the test slams on the brakes. Sometimes with the forward collision warning and.. it's like 200 meters ahead of you.. like they'll easily turn out past you.. But.. rear ending potential.. that's the worry. (R006 - F/SCW)"

- **Inadequate distance - ad-SOTIF(+) PST(-)**
  Finally, one driver reported that the vehicle failed to maintain a sufficiently safe distance from a parked vehicle, even though it did not result in a collision. This situation made

[1] Blind Spot Monitoring (BSM) was not included in this category, as the only driver who expressed low trust in the BSM system attributed this to the placement of the warning symbol rather than concerns about its functional performance.

them feel unsafe:

"I'm not gonna say terribly unsafe, but uncomfortable. I do feel very unsafe if there's vehicles parked on the right hand side and the the vehicles attempting to maintain the lane, but it comes far too close to the vehicles on the right hand side. That that is very I feel that's very unsafe and that's that's very stressed. (R041 - LDA)

- **Inappropriate braking - ad-SOTIF(+) PST(-)**
  The driver described experiences in which the vehicle's emergency braking behaviour was problematic, particularly due to hesitation after braking in heavy traffic. While the system still performed effectively in preventing collisions, the driver implicitly expressed concerns about safety due to this behaviour. Specifically, if traffic clears and speeds up, such behaviour could disrupt the flow of traffic, as it acts in a way that is not anticipated by other drivers.

  "The emergency brake checking that goes on, where the car will break.. in heavy traffic. It's okay when the car's hitting the brakes and hesitating. But when it opens up, and we're moving faster, and there's more space.. People are anticipating you to stay at your speed.. You don't want.. hitting the brakes at those speeds.. those are the biggest situations. (R081 - AEB)"

Although AEB, F/SCW, and LDA successfully tracked vehicle manufacturers' safety expectations—such as braking to prevent collisions in heavy traffic, issuing warnings of potential collisions, and maintaining lane position—drivers reported several concerns that negatively affected their perceptions of safety and trust. These issues suggest that the active safety features did not consistently align with drivers' safety expectations. For example, AEB was reported to brake unnecessarily or hesitate in dense traffic conditions, disrupting traffic flow and raising safety concerns. F/SCW was criticised for issuing premature warnings and engaging in unnecessary braking when no imminent threat was present, often leading to frustration. LDA was noted for failing to maintain a safe lateral distance from parked vehicles, which resulted in driver discomfort and a diminished sense of security.

Consistent Tracking

Two active safety features—BSM and LDA—were categorised as exhibiting consistent tracking, as they were reported to effectively meet both vehicle manufacturers' and drivers' safety expectations in most scenarios.

The specific situations in which drivers indicated that LDA successfully aligned with their safety expectations are outlined below.

- **Long trips - ad-SOTIF(+) PST(+)**
  Two drivers highlighted how their vehicle performed exceptionally well in maintaining its lane during long trips, particularly on highways. They described the feature as "flawless," suggesting that they perceived the system as both safe and trustworthy.

  I did a.. 7500 Mile road trip from Connecticut to California and back.. 99% of the

trip on the highways.. was done using Autopilot. And it worked pretty much flawless. (R026 - LDA)

Additionally, one driver noted that the system outperformed human drivers, particularly in maintaining focus and avoiding complacency during extended drives. This suggests that the driver trusted the system to remain vigilant and avoid complacency.

For autopilot.. used on a highway, I would say I'm the worst driver in the fact that it does a better job of long distance drives keeping lane centered. You know, watching.. not becoming complacent, I guess, which is so easy on on a longer drive. (R087 - LDA)

- **Managing complex highway infrastructure - ad-SOTIF(+) PST(+)**
  Two drivers highlighted how their vehicle assisted them in navigating complex highway traffic. One driver emphasised that FSD Beta maintained their lane and did not drift into incoming on-ramps, noting that it performed merging manoeuvres better than Autopilot.

  There's been a few highways where FSD beta can be engaged at highway speeds. And it does solve many of the problems they've had with navigate on autopilot, and then it merges better. It doesn't shift over into incoming on ramps like navigating like I'll it does it steers better. It maintains speed better overall. (R007 - LDA)

  Another driver described a highway they considered a "scary" place to drive due to the numerous on- and off-ramps on both sides of the interstate. They noted that Autopilot kept them in the correct lane, something they felt they might not have been able to maintain on their own.

  I don't think I could get through Atlanta if I didn't have Autopilot because their interstate is twice as wide as ours and they have on ramps and off ramps on both sides of the interstate. And it is a crazy, hectic, scary place to drive and. If I didn't have autopilot keeping me where I needed to be, I don't think I could do it. Nerves of steel there and I don't have it. (R099 - LDA)

- **Less mental workload - ad-SOTIF(+) PST(+)**
  Drivers consistently reported experiencing a reduced mental workload when using the vehicle's lane-keeping features. This reduction in cognitive effort was attributed to the vehicle's ability to handle routine tasks, such as maintaining lane position and adjusting for nearby traffic. One driver described how the system's reliability in keeping the car centred in the lane fostered a sense of security, allowing them to relax and trust the technology:

  I trust full self-driving to keep me in my lane. So no, I don't pay as close attention to where I am in the lane. I trust that it's keeping me in the lane. (R044 - LDA)

  One driver emphasised that the system's effectiveness in maintaining lane position significantly reduced fatigue, leading to a more positive driving experience.

And the reason it makes it a lot less fatigue.. is that you don't have to mentally think about all the micro adjustments. So when you're driving down the road, you have to constantly make sure you're centered in the lane, make sure you're keeping distance from the car in front of you.. That's my experience.. I really positive with Autopilot now for Full Self Driving Beta. (R079 - LDA)

Another driver described how the system allowed them to shift their focus towards broader situational awareness, which they considered a safer and more efficient way of driving:

I'm no longer having to concentrate on keeping that car.. I just simply don't even think about it anymore. In fact, it's odd when I take it off of all the pilot effect man, this is like starting out driving again all over because it's just something you get used to that the car keeps it so well on its lane that you just don't think about that anymore. What you do is looking ahead. You're looking for other things happening to you and you're making sure that you react appropriately. Does really good. (R062 - LDA)

Despite the reported excellence of LDA, one driver highlighted a situation where LDA kept them in the wrong lane, which led to them feeling scared. While LDA was still functioning as intended by keeping the vehicle within the lane, it did so in the wrong lane.

- **Stay on the wrong lane - ad-SOTIF(-) PST(-)**

  "It's really scary. It just does all sorts of weird things today. I was like coming home from work and it stayed. It was two lane road. It stayed in the left lane, which turned into a turn lane and it just like blew right through the turn lane and just kept writing through. We call it here as suicide lanes where you have a you can make a left or a right turn either direction. And it just kept driving right through it. (R076 - LDA)"

The following aspects were highlighted regarding the effectiveness of the BSM system.

- **Safe lane changing - ad-SOTIF(+) PST(+)**
  Drivers consistently praised the vehicle's BSM system as a crucial safety feature that enhances the driving experience during lane changes. One participant highlighted that the system effectively monitors the vehicle's surroundings and facilitates safer lane changes by detecting vehicles in blind spots. This feature was reported to significantly increase their confidence and sense of safety while manoeuvring between lanes.

  A very complete functionality, features and.. ability to.. monitor everything around you and that lets you change lanes if there's a car in your blind spot or coming through and you're using it and stuff like that. Definitely makes me feel much safer when I'm doing it. (R026 - BSM)

  Another participant expressed appreciation for the BSM feature, noting that it helped prevent accidental lane changes resulting from limited peripheral vision. They found the BSM display particularly useful for enhancing situational awareness and reducing the likelihood of unintended lane merges.

When we go on vacation.. we're gonna be doing a lot of miles.. driving across the country. It takes a lot of.. drive.. I.. really enjoy it because I am blind on my entire right side. I have no peripheral vision, so it makes it with the screen being there and you know, blind spot awareness and all of those interesting features. It makes it harder for me to accidentally merge into someone if I don't look forward enough to the side to see if anybody's in there. (R099 - BSM)

- **Understanding what the system perceives - ad-SOTIF(+) PST(+)**
  Drivers reported that the Blind Spot Monitoring (BSM) system enhances their awareness of the vehicle's surroundings. One participant expressed appreciation for the system's graphical display, which allowed them to compare the vehicle's sensor feedback with their own visual observations. This feature was perceived as highly accurate and contributed to a greater sense of situational awareness.

  I would say.. 90% of the time my eyes are on the road. You typically monitor vehicle and its surroundings at all times.. I also enjoy the graphic that it gives you so you can understand our like to constantly compare with the vehicle sees to what I see and see what I can spot. That vehicle doesn't yet. And for the most part it's. Like 95% accurate. (R032 - BSM)

  Another participant emphasised that the visual feedback provided by the system on the display screen enhanced their confidence, as it allowed them to see exactly what the vehicle was detecting. This level of transparency contributed to a greater sense of safety and environmental awareness, reinforcing their trust in the system's ability to identify and avoid potential hazards.

  For both of them it's. You know, I feel safer because I see the perception. On the screen so I can see what it sees. And you know that gives me confidence of. Knowing exactly what what it is seeing compared to.... And a lot of the perception part of this avoids that. Avoid those situations or helps avoid the situations. (R087 - BSM)

However, one participant reported a case in which the BSM system did not function as intended, resulting in an unsafe situation. According to the driver, the malfunction was caused by direct sunlight interfering with the sensor's ability to detect surrounding vehicles.

- **Weather-related sensor limitations - ad-SOTIF(-) PST(-)**
  "The place where I feel it's starting to get unsafe is the changing weather conditions. And sometimes lighting. That's one other one. When you get a a bright hit of sunlight across into one of the panel doors, it'll just blind the camera. It can't compensate, and some levels. And I think they're gonna have to improve some of the cameras all around the car to be able to decrease their contrast to avoid it. These are the situation with you so unsafe. (R061 - BSM)"

Both LDA and BSM were noted for effectively tracking both vehicle manufacturers' and drivers' safety expectations. Specifically, LDA was praised for its ability to maintain lane position during extended highway travel and in complex driving environments, contributing to

reduced driver fatigue and mental workload. BSM was commended for enhancing safety and driver confidence during lane changes by reliably monitoring blind spots and improving overall situational awareness. This feature was particularly valued by drivers with limited peripheral vision, who found the system especially beneficial.

Despite these strengths, instances of tracking failures were reported. BSM occasionally failed to function correctly due to sensor interference from direct sunlight. In the case of LDA, one participant reported a failure to maintain lane position, although the precise cause of this issue could not be determined.

Table 6.3: Assessment of the tracking condition of meaningful human control based on common situations mentioned by users of partially automated driving systems. A positive mark (+) indicates that the respective human agent's expectations are tracked, while a negative mark (–) indicates that the expectations are not tracked. The final column indicates whether both the driver's and the automaker's expectations are tracked.

Safety Feature	**Described situation**	**Tracking of driver's expectations (PST)**	**Tracking of automaker's expectations (ad-SOTIF)**	**Human expectations are ..**
BSM	Driver can detect objects in their blind spot while driving	+	+	Tracked
	BSM's sensors dysfunction due to weather such as sunlight	–	–	Not tracked
LDA	Driving on long highway trips with complex driving conditions	+	+	Tracked
	Drivers don't have to perform minor adjustments of the vehicle within its lane	+	+	Tracked
	LDA keeps the vehicle on the wrong lane	–	–	Not tracked
	LDA maintains lane, but the distance with surrounding objects is too close	–	+	Partially tracked
F/SCW	F/SCW warns the driver of unseen potential front and side collisions with sound and on-screen icons	+	+	Tracked
	F/SCW warns the driver of potential collisions that are still distant	–	+	Partially tracked
	F/SCW responds to false positive information and does not react to false negatives	–	–	Not tracked
	Annoying warnings after system reboots	–	–	Not tracked
	Warning dysfunctions when vehicle with high speed approaches	–	–	Not tracked
AEB	AEB brakes to prevent collision in unforeseen/unexpected situations	+	+	Tracked
	AEB responds to false positive information and does not react to false negatives	–	–	Not tracked
	AEB brakes to prevent collision, but the driver dislikes the way it brakes	–	+	Partially tracked

Summary

To summarise the tracking evaluation results, we aggregated the commonly reported situations for each safety feature across the three categories discussed in Sections 6.4.1 to 6.4.1 (Table 6.3). The classification of 'tracked' or 'not tracked' is based on the presence of recurring themes in participant responses for each safety feature, as described in the corresponding sections. If a particular theme was mentioned by participants, the safety feature was classified accordingly in the table.

This analysis revealed that for each safety feature, there are situations in which the feature successfully tracked both the driver's and the vehicle manufacturer's safety expectations, as well as situations where it did not. Notably, failures to track the vehicle manufacturer's safety expectations were always accompanied by failures to track the driver's safety expectations. However, the reverse was not always true; in some cases, the system met the manufacturer's expectations but failed to align with the driver's expectations.

6.4.2 Tracing Evaluation Results

Using thematic analysis, we evaluated tracing by identifying ten subcategories within participants' responses, corresponding to the four tracing evaluation criteria. We also analysed the number of drivers who mentioned each subcategory (Table 6.4). These subcategories provide insight into how drivers operationalise the tracing criteria in practice, offering a deeper understanding of how responsibility, knowledge, and control are perceived and enacted.

Table 6.4: Sub-categories related to tracing evaluation criteria. For each sub-category, count indicates the number of participants who mentioned each sub-category.

Tracing evaluation criteria	Sub-categories	Count
Knowledge: keeping both hands on the steering wheel	Driving with both hands on the steering wheel	39
	Driving with one hand on the steering wheel	13
	Driving mode	16
Knowledge: staying alert	Observation of the surrounding situations	17
	Highway	7
	Driving mode	26
Capacity: corrective action	Control over steering wheel	19
	Control over the pedals	10
	Driving mode	28
Maintaining operational responsibility	Agree to maintain control and responsibility	19

The frequency of mentions also indicates that certain subcategories were discussed more frequently than others. For instance, the 39 references to driving with both hands on the steering wheel suggest that a relatively large number of participants either understood or actively

practised this behaviour. This number is notably higher than the 13 mentions of driving with only one hand on the wheel.

To provide deeper insight, the following sections offer detailed explanations and representative quotations for each tracing evaluation criterion, along with their corresponding subcategories.

Knowledge: Keeping Both Hands on the Steering Wheel

This tracing requirement concerns whether drivers possess adequate knowledge regarding system use. Specifically, we assessed whether participants understood the importance of keeping both hands on the steering wheel and whether they reported complying with this guideline. Based on the interview data, we identified three categories of responses that addressed this topic.

- **Driving with both hands on the steering wheel:** Several participants reported consistently keeping both hands on the steering wheel. This behaviour was often attributed to legal requirements, with some noting that they adhered to this practice to avoid reprimands or penalties.

  "Do you typically keep your hands on the steering wheel at all times? I do (R041)"

  "According to the law, the hands must be on the wheel. I actually keep my hands on the wheel, and I feel the resistance. (R047)"

  "I do keep my hands on the steering wheel mostly so I don't get dinged. (R067)"

- **Driving with one hand on the steering wheel:** Other participants reported typically keeping only one hand on the steering wheel. For some, this was primarily to meet the system's torque detection requirements, while others adopted this behaviour when using Autopilot, often describing a more relaxed driving posture during such instances.

  "I typically keep one hand on the steering wheel at all times. I keep it there just enough to satisfy the torquing requirement, where there needs to be weight on the system." (R054)

  "Do you always keep both hands on the wheel? No, generally, I keep one hand. So, I have my left hand always on the wheel. It's usually on my knee, on the door, or on my elbow.(R087)"

  "Yes, I keep my hands a little bit stream at all times. When.. I'm not have my hands on the steering wheel, I either have one hand like on the bottom like like like one hand in this picture. But I at least always have one hand on the steering wheel. (R098)"

- **Driving mode:** Participants reported varied behaviours concerning hand placement on the steering wheel depending on the driving mode. While some consistently used one hand when operating Autopilot, others indicated that they were more likely to keep both hands on the wheel when using FSD Beta. The responses also revealed a range of strategies for maintaining system engagement while using Autopilot, including resting hands

underneath the wheel, intermittently jiggling it to satisfy system prompts, or applying continuous pressure with one hand to meet torque detection requirements.

"I'll usually have just one hand.. just lean my hand on the bottom of the steering wheel and let the weight kind of be enough to do it, so that's generally how I drive with just to put enough pressure on it. Keep it constant pressure on it so it never really warns me about not putting pressure on. I tried to do things around, just occasionally do it, but that becomes more effort than just letting your hand rest on the steering wheel when I'm driving with it.. We usually just keep my hand sitting there resting there and it works.(R021)"

"Then do you typically keep your hands on the steering wheel at all time? Autopilot no, FSD beta yes (R033)"

"With Autopilot.. depends on where we're.. if we're on the highway.. where there are no obvious issues up ahead that I can see, what I'll typically do is rest my hands underneath the wheel. And then as the prompts come up, I'll just jiggle the wheel a little bit to make the prompt go away. With.. beta, most of the time.. Especially during turns. Typically.. I'll have my hands at.. seven and four or something, and just let the wheel sort of brush up against my hands. And sometimes I'll keep my hands off the wheel.. if I'm comfortable in this situation. But I've kind of learned not to do that. (R051)"

Overall, the evaluation of drivers' knowledge regarding the requirement to keep both hands on the steering wheel revealed a range of practices. While some participants consistently used both hands in adherence to legal requirements and to avoid penalties, others adopted a more relaxed approach—maintaining one hand on the wheel primarily to satisfy the system's torque detection, particularly when using Autopilot. Behaviours also varied by driving mode; drivers were generally more likely to maintain a hands-on approach when using FSD Beta compared to Autopilot. From a tracing perspective, although drivers appeared to understand the requirement to keep their hands on the steering wheel, their actual behaviours demonstrated considerable variation.

Knowledge: Staying Alert

According to the vehicle manufacturer, drivers are required to remain alert at all times. Based on this requirement, we identified three subcategories of participant responses:

- **Awareness of surrounding situations:** Participants expressed varied perspectives regarding situational awareness. Some reported that using Autopilot and FSD Beta enhanced their attentiveness, allowing them to focus further down the road and experience reduced fatigue. One participant noted experiencing heightened alertness while using the technology, citing improved contextual awareness and the ability to more effectively scan the driving environment. Another emphasised the importance of remaining fully attentive, particularly when using the beta version, to stay aware of the surrounding conditions.

  "Autopilot and FSD beta allow you to actually be more attentive in general then not having autopilot or FSD, and that's because the car is taking care of the rudimentary

things for you.. That allows you to focus further down on the road or it allows you to see things that maybe you wouldn't have seen otherwise and it allows you to be less fatigued to where you're able to. You're able to be more alert than you would be otherwise. That doesn't mean that you don't also get distracted at times, but I think when you are paying attention, I think it allows you to pay better attention to the road than than without autopilot or FSD. (R027)"

"For Autopilot and FSD always be alert and attentive. If it's a beta, it's required to be fully attentive and alert at all times. Autopilot. I know other owners, they're kind of relaxing, not paying attention. For me on Autopilot, it helps me become more attentive of my surroundings during driving.. When I'm driving myself, I usually look forward in once in a while, look left and right, but with autopilot, I'm able to watch.. all the mirrors all the time, making sure what I'm aware of what's going on around me. (R075)"

"Typically, fully attentive and alert at all times. Pretty excessively alert. As one of the things I love about the the beta and regular autopilot as well, when they drive it, actually more aware because I can actually look around and take in.. where all the cars are around me. I understand.. what's going on, where it was.. I definitely enjoy it more when I'm not micromanaging those things and I'm able to take in and be more contextually aware. (R085)"

- **Highway:** Participants' experiences with staying alert while driving varied depending on the driving context. In particular, some reported reduced attentiveness when using Autopilot on highways—especially on familiar routes, during low-traffic conditions, or in the absence of external distractions. The following quotations illustrate this variability in driver alertness:

  "And typically fully attentive and alert, you have to be.. Autopilot ..not so much.. I've noticed.. I'll be driving along and.. be able to read a sign along the road or something that. Before, you.. you wouldn't take the time to read and add. But.. if it's on a Interstate highway, it's no problem. (R010)"

  "I typically fully attentive and alert at all times? No, I've gotten comfortable with it over time, so I don't fully pay a attention anymore, specially on roads that I'm familiar with or on highways that I'm familiar with. (R016)"

  " I would say there are moments.. where I haven't been fully attentive. I obviously don't let that happen for like minutes.. I'm not gonna pull my phone out, look at it, but there's definitely times when driving on the highway, I look ahead and there's nobody for a kilometer ahead of me. And so I will look after the side and look at something in the scenery and then look back again or look at the passenger beside me and then look back again. Not for long periods of time. But longer than you could get away with if you were actually the one driving, I would say. (R045)"

- **Driving mode:** Participants reported variations in their levels of attentiveness and alertness depending on the driving mode. One participant noted that their awareness was lower when using Autopilot compared to FSD Beta, even when feeling fatigued. An-

other acknowledged being less attentive in Autopilot mode, explaining that it enabled multitasking behaviours that would not be possible during conventional driving. A third participant stated that their level of attentiveness while using Autopilot was slightly lower than when using FSD Beta. The following quotations offer further insight into these reported differences:

"Are you typically fully attentive and alert at all times? I'd say with Autopilot I have been in situations where I've driven really exhausted and I tend to have pretty good situational awareness even when I'm.. super exhausted. But I would say like.. when I use autopilot, it's not always 100 percent peak performance.. With FSD beta.. I'm always alert and fully attentive.(R051)"

".. If I were to grade these on on how I feel on where I have to be fully attentive and alert at all times, FSD beta requires the most, Autopilot requires less(R063)"

"..typically fully attentive, fully alert.. at all times.. Less.. in autopilot.. I would say that Autopilot does allow you to do other things that you might not normally do if you were driving the car normally. (R088)"

The evaluation of drivers' knowledge regarding the need to remain alert while driving revealed three key insights in relation to the tracing condition. First, some participants reported that Autopilot and FSD Beta enhanced their attentiveness by allowing them to focus further ahead and reduce fatigue. Second, attentiveness varied depending on the driving context; some drivers reported decreased focus on highways, particularly on familiar routes. Third, alertness differed across driving modes, with participants generally reporting lower levels of awareness while using Autopilot compared to FSD Beta.

Overall, although participants demonstrated knowledge of the requirement to stay alert—thus formally satisfying the tracing condition—their actual behaviours reflected variability in alertness depending on context and system use.

Capacity: Corrective Action

The tracing condition of MHC requires that drivers not only understand the functionality of partially automated systems but also retain the capacity to operate them effectively. Consistent with this requirement, the vehicle manufacturer in our study mandates that drivers must be able to perform corrective actions. Based on our analysis of participant responses, we identified three sub-categories reflecting this capacity.

- **Control over the steering wheel:** Participants demonstrated readiness to take corrective action through their hand positioning while using automated driving features. One participant emphasised maintaining a firm grip on the steering wheel, deliberately placing their hands in the lower corner to enable a rapid response to unexpected lane drift. Another respondent noted that their approach to hand positioning was influenced by how Autopilot handled specific driving situations. Additionally, one participant reported that resting one hand on the wheel was ineffective for performing minor corrective actions when using

Autopilot.

"Often when I'm driving on the highway.. if it's just me, I'll just have one hand resting on top of the wheel, making minor corrective action. But that doesn't work very well with Autopilot. It thinks that I'm not touching my car so. So I yeah, do like the nine and three or five and seven to use gravity. (R028)"

"Do I prepare to take corrective actions? Absolutely, whether it's holding that steering wheel really hard in case it wants to just drift off really quick or.. really hard.. that's.. why I hold my hands. The way they're steering wheels made, that's also why I hold my hands in that lower corner as opposed to up top when you hold it up top. If the car is gonna jerk itself off to the right, especially being left handed, it can only go so far before the center beam and the steering wheel will block it. But if you hold it on the bottoms, it has much less traveled before you can get your hand on it. And if anything else, it's going to stop as soon as it hits you hit one of that center peg.(R061)"

".. Stay prepared to take corrective actions, like more.. than if I was just tracking upon myself. Because technically there's someone else driving the.. car, you know that they're not very good at driving the car, so I have to pay more attention.. I guess I'm pretty good spatial awareness, so I take a lot.. for granted.. in terms how you place your hands on the steering wheel. (R081)"

- **Control over the pedals:** Participants described their interactions with the brake pedal as a means of demonstrating their readiness to perform corrective actions. One participant noted the importance of being prepared to intervene, particularly in situations where FSD or Autopilot might fail to brake in time. Another explained that they positioned their foot in a comfortable location to enable rapid braking or acceleration when necessary. A third participant emphasised the ease with which they could engage the brake, as their foot was already positioned similarly to when operating vehicles with lower levels of automation.

  "Do you typically stay prepared to make corrective actions at all time? Absolutely, especially with FSD, better you have to be prepared. You have to kind of.. exit plan. If it comes down to it with Autopilot, not as much. But there are times where you may have to be ready to press the brakes because the car is not breaking in time and it's getting a little bit too close to the car ahead of you. (R064)"

  "I always stay prepared to take corrective action with FSD beta. But with autopilot on the highway.. feel a little bit more comfortable with my foot. Like to the side.. But it is really easy to lift my foot and hit the brake if I need to.. I would say like it's pretty much the same as when I use cruise control on older cars or other cars in the past. It's the same place. I would put my foot.(R074)"

  "Do you typically stay prepared to take corrective actions at all times? Mostly I keep my feet just like back. I'll wait from the pedals. Just getting a comfortable position unless.. a location where I don't have as much trust in Autopilot, FSD beta, or there's a lot of cars around me. Then I have my feet ready to like, break or accelerate or anything like that mostly depends on the situation. (R078)"

- **Driving mode:** Participants' perspectives on their preparedness to take corrective action varied depending on the driving mode. One respondent strongly emphasised the importance of maintaining vigilance while using FSD Beta, describing a constant state of readiness to intervene. Another participant highlighted a contrast between the two modes, reporting a heightened level of awareness—described as being "hyper-aware"—when using FSD Beta, and a lower level of preparedness when using Autopilot. Additionally, some participants indicated that their trust in Autopilot increased over time, leading to a more relaxed driving posture and a perception of the system as being safer. The following quotations illustrate these perspectives:

  "Do you typically stay prepared to take corrective action at all times? Autopilots a little bit less than FSD beta. As long as I'm in a comfortable realm, there's no situations around me. I am prepared, but my guards down a little bit more. With FSD beta. I'm always ready to take over. (R017)"

  "Do you typically stay prepared to take corrective actions at all times? I certainly do that.. for all the reasons.. But autopilot, I feel like I'm typically less prepared because I'm more relaxed. I'm more letting my guard down. Because.. I trust it more. It's never done as much wrong as.. I'm looking at the scenery I'm looking.. I'm enjoying the ride versus driving pretty much. So definitely less prepared on autopilot.. Autopilot is safer in my opinion. (R048)"

  "Next question that typically stay prepared to take corrective action. Autopilot.. not so much. I mean, my hand is on the wheel.. With Beta. I'm very.. ready to take control. It's hyper aware. (R096)"

The evaluation of drivers' capacity to perform corrective actions—another key component of the tracing condition—revealed that participants demonstrated readiness by adjusting their grip on the steering wheel and maintaining their foot near the brake pedal. Several drivers also adopted specific strategies to enable rapid intervention when necessary. Preparedness varied by driving mode: participants reported feeling more vigilant while using FSD Beta and more relaxed when using Autopilot. Overall, the findings suggest that drivers exhibited different levels of readiness to take corrective action, influenced by both the automation mode and their individual driving strategies.

Maintaining Operational Responsibility

The final tracing requirement concerns the maintenance of operational responsibility. According to the tracing condition, at least one human must be aware that they hold moral responsibility for the outcomes of the system's actions. This aligns with the vehicle manufacturer's guidance, which stipulates that drivers are accountable for the operation of the partially automated driving systems.

Unlike the other tracing evaluation criteria, only one subcategory was identified in this area: agreement among participants that drivers must maintain control and assume responsibility.

Participants expressed this recognition in varying ways. One respondent mentioned feeling

personally responsible for ensuring that the vehicle did not make errors. Another indicated a strong belief that, in the event of an incident, they would be held fully liable and could not shift blame to Autopilot in a legal context. A third respondent explicitly emphasised the importance of maintaining control and responsibility, acknowledging that they would accept fault in the case of a collision. The following quotations provide illustrative examples:

"The responsibility is definitely mine,. I wrecked my car.. not tesla fault.. indeed. (R032)"

Did not maintain control in this? No, I disagree with that. I mean..I get that I'm completely responsible for it. I'm gonna lose in court if I say Autopilot made me did it, or autopilot did it.(R067)

"I'm paying attention to what it's doing, backing it up to make sure it doesn't make a mistake.. But.. if it does, I'm responsible for it. So I have to be really paying attention to it. So I'm vigilant. But.. I feel like probably secure that it's doing a good job. (R072)"

Overall, the evaluation indicates that drivers are aware of their responsibility to oversee the vehicle's operation and recognise their accountability in the event of system errors or legal consequences.

6.5 Discussion

6.5.1 Theoretical Implications

This section discusses how the proposed MHC evaluation framework can be applied to systems based on real-world driving experiences, offering new insights into the dynamic nature of meaningful human control (MHC), particularly in relation to the tracking and tracing components. The findings highlight the interplay between system performance and human factors, contributing to the existing body of literature by emphasising the roles of contextual variability, subjective risk perception, and the interaction between human engagement and system behaviour in the assessment of MHC.

The tracking evaluation revealed notable variations in how different safety features align with both human- and manufacturer-defined safety expectations across varying driving contexts. Features such as Blind Spot Monitoring (BSM) and Lane Departure Avoidance (LDA) demonstrated strong alignment with the tracking component of MHC during routine scenarios, such as highway lane-keeping. Drivers particularly valued BSM's warning system and visual interface for identifying vehicles in blind spots, consistent with findings from Kim et al. (2024), who reported that user interfaces offering surrounding information enhance driver trust and reduce perceived risk.

However, in emergency or unexpected driving situations—such as encounters with sudden obstacles—features like Automatic Emergency Braking (AEB) and Forward/Side Collision Warning (F/SCW) exhibited performance inconsistencies. Although drivers acknowledged their effectiveness in preventing collisions, these systems were less reliable in consistently meeting the tracking requirements of MHC. This observation aligns with Cicchino (2017), who found

that such features significantly reduce front-to-rear crash rates but are not universally effective. These results underscore the importance of ensuring that partially automated systems are capable of dynamically adapting to diverse and unpredictable driving environments in order to uphold meaningful human control.

Misalignment between driver and manufacturer safety expectations often arises from technological limitations—such as sensor failures in adverse weather conditions, including bright light impairing sensor performance—or from mismatched expectations, such as drivers perceiving AEB braking as overly hesitant. Tesla's manual (Tesla, 2024b) explicitly acknowledges limitations such as obscured lane markings or weather-related interference (e.g., rain); nevertheless, drivers expressed safety concerns when these limitations manifested in practice. These findings underscore the importance of addressing root causes, including the enhancement of sensor reliability and better alignment of system behaviour with human expectations.

In addition, drivers' risk perception played a critical role in the tracking component of MHC. False positives, such as phantom braking, diminished trust in the system, whereas successful interventions, such as timely collision avoidance, improved perceived safety. This dynamic highlights the need for the tracking component of MHC to account for both objective system performance and the subjective experiences of human drivers.

The tracing evaluation reveals how human factors shape the effectiveness of partially automated systems in meeting tracing criteria. Drivers frequently engaged selectively with system warnings or interventions based on their personal risk assessments. For instance, some participants reported disregarding Forward/Side Collision Warning (F/SCW) alerts when they judged the following vehicle to be at a safe distance, indicating a disconnect between system logic and human judgement.

The paradox of trust also emerged as a critical influence on tracing compliance. While drivers expressed appreciation for features like LDA for reducing cognitive load—consistent with findings by Miller & Boyle (2019), who demonstrated increased workload in the absence of LDA—over-reliance on such features often resulted in complacency. This supports the argument of Bainbridge (1983), who described the "ironies of automation," wherein human vigilance diminishes as system reliability increases. Young & Stanton (2002) further conceptualise this effect through "mental underload," where reduced task demands lower attentional capacity and compromise readiness to intervene.

Although manufacturers attempt to mitigate this risk by assigning drivers the responsibility to remain engaged, in practice, drivers often disengage during routine operation, becoming "out-of-the-loop" (Endsley, 2017). This challenge is exacerbated by system design approaches that overlook human cognitive limitations in sustained attention and monitoring tasks (Lee & See, 2004). The resulting paradox exposes a fundamental design flaw: assigning moral responsibility alone is insufficient to guarantee continuous vigilance. As argued by Hansson et al. (2021), systems that promote over-reliance while simultaneously expecting uninterrupted human supervision raise significant ethical concerns.

The level of driver engagement was found to vary depending on system behaviour and driving context. Participants tended to be more engaged in complex or high-risk driving scenarios, while disengagement was more common during routine tasks. This dynamic aligns with findings by Robins-Early (2024) and **?**, who demonstrated that subjective risk perceptions—such as preferences for lateral distance during automated overtaking—significantly influence trust and

perceived safety. For instance, some drivers reported experiencing stress when LDA maintained a minimal lateral buffer, even if the manoeuvre was technically safe.

Personal preferences, driving modes, and situational contexts also played a significant role in tracing performance. Drivers were more likely to comply with hand placement requirements in urban environments when using FSD Beta—perceived as riskier—while adopting minimal contact strategies (e.g., resting a hand or applying intermittent torque) on highways with Autopilot, which was perceived as more stable. Additionally, some participants admitted to deliberately manipulating the system by applying weight to the steering wheel to simulate compliance with hand detection requirements.

This pattern of selective adherence highlights the complex interplay between individual attitudes, perceived risk, and contextual factors. It suggests that tracing performance cannot be fully understood without accounting for how drivers interpret and respond to system cues within specific driving environments.

6.5.2 Practical Implications

This section offers practical recommendations based on the insights obtained from the MHC evaluation, with a focus on enhancing system design and addressing subjective driver experiences.

To improve system design, it is essential to address environmental limitations such as glare from sunlight, adverse weather conditions, and faded lane markings, all of which can impair system reliability. For example, Blind Spot Monitoring (BSM) sensors that are susceptible to sunlight interference could be redesigned using alternative sensing technologies or with added redundancy to ensure consistent performance. Minimising false positives and false negatives—such as phantom braking or missed hazard detections—is also critical for sustaining driver trust. Potential solutions include refining object detection algorithms and integrating contextual awareness to reduce unnecessary alerts. These design improvements are urgent, particularly in light of findings by Paula et al. (2023), who reported that 78% of drivers were unable to override phantom braking, thereby heightening safety risks.

Moreover, system behaviour should be calibrated to align more closely with human expectations of safety and trust. For instance, Automatic Emergency Braking (AEB) could be adjusted to engage earlier in emergency scenarios, reflecting drivers' preferences for proactive intervention. This recommendation is supported by Koglbauer et al. (2018), who demonstrated that braking behaviour significantly influences perceived safety.

Addressing subjective driver experiences is equally critical for improving the effectiveness of partially automated systems. Clear and intuitive user interfaces can enhance driver understanding of system behaviour. For example, visual or auditory cues explaining why a warning was issued or why braking occurred could reduce confusion and strengthen trust. Encouraging driver engagement is also essential. Systems should actively prompt drivers to assume control in edge cases—such as when lane markings are unclear—to mitigate over-reliance on automation. Furthermore, educating drivers about system capabilities and limitations remains a key priority. Comprehensive training programmes can support more effective use of automation by emphasising the importance of staying attentive and prepared to intervene.

To address the broader challenges of driver engagement and over-reliance on automation, we propose several actionable recommendations. First, adaptive Human-Machine Interfaces (HMIs) could be developed to tailor hand placement reminders based on the driving context. For instance, stricter prompts may be warranted on highways, where automation is typically perceived as reliable, whereas fewer prompts may be appropriate in urban settings, where drivers are more naturally engaged due to increased perceived risk.

Second, enhanced training and regulatory measures are essential to reinforce driver readiness. Scenario-based training modules could prepare drivers to respond effectively in low-risk contexts, while policies mandating multi-layered engagement checks—extending beyond easily circumvented measures such as steering torque detection—would promote sustained vigilance.

Finally, reassessing preparedness expectations is crucial. Given the natural constraints of human attention and the observed tendency to over-rely on automation, vehicle manufacturers should reconsider assumptions about how quickly and effectively drivers can retake control. By incorporating these recommendations, automated driving systems can better align with human capabilities and limitations, thereby ensuring that drivers remain meaningfully engaged and ready to intervene when necessary.

6.5.3 Are Tesla's Partially Automated Driving Systems Under Meaningful Human Control?

Based on our evaluation of **tracking** and **tracing compliance**, we conclude that Tesla's FSD Beta and Autopilot systems do not fully satisfy the requirements of meaningful human control (MHC). In contrast to previous studies that primarily assess MHC through hypothetical scenarios or post-incident analyses (Calvert et al., 2021, 2020b), our evaluation provides a more nuanced understanding of real-world system behaviour and its implications for MHC compliance. Below, we summarise the key findings that support this conclusion.

Failures in Tracking Compliance

Tesla's systems frequently failed to track safety expectations under challenging environmental conditions and in the presence of degraded infrastructure. These shortcomings underscore a lack of robustness in the perception systems, which are essential for maintaining alignment with both driver and manufacturer safety expectations. For example, adverse weather conditions—such as rain, snow, or glare—were reported to impair sensor functionality, resulting in failures to detect obstacles or maintain appropriate lane positioning. Similarly, infrastructure-related issues, including faded lane markings and poorly maintained roads, further exacerbated these limitations, as the systems rely heavily on visual inputs for accurate operation.

In high-risk or unpredictable scenarios, features such as Automatic Emergency Braking (AEB) and Forward/Side Collision Warning (F/SCW) often struggled to effectively track safety expectations. Issues such as phantom braking—where the system erroneously detects obstacles and applies the brakes unnecessarily—and false negatives—where genuine hazards go undetected—further compromised performance. These inconsistencies not only eroded driver trust but also diminished the system's capacity to meet safety expectations in critical situations.

Moreover, even when systems conformed to technical specifications, such as adhering to predefined braking thresholds, they frequently failed to meet driver expectations. Subjective perceptions of safety often diverged from objective system performance. For instance, participants described AEB interventions as overly cautious or hesitant, despite the system functioning within its intended parameters.

Failures in Tracing Compliance

Failures in tracing compliance stem from inconsistent driver adherence to safety protocols, over-reliance on automation in low-risk scenarios, and systemic design shortcomings that inadvertently promote disengagement. Drivers frequently demonstrated selective adherence to recommended behaviours, such as maintaining hands on the steering wheel and staying alert. Higher compliance was observed in high-risk contexts—such as urban environments with FSD Beta—where the perceived complexity of the driving environment prompted greater vigilance. Conversely, in low-risk contexts such as highway driving with Autopilot, compliance levels declined substantially as drivers placed greater trust in the system's reliability.

This variability indicates a troubling dependence on driver confidence rather than on system robustness to ensure safe operation. It also reveals a fundamental challenge in tracing compliance: current systems often fail to support sustained driver engagement and accountability, especially during routine or low-demand driving conditions.

An inverse relationship was observed between driver confidence and preparedness to perform corrective actions. When drivers perceived the system to be safe—such as during routine highway driving with Autopilot—their vigilance and readiness to intervene declined. This over-reliance on automation introduces significant risk, as drivers may be insufficiently prepared to take control in emergency situations, thereby undermining the system's ability to maintain meaningful human control.

Although participants generally acknowledged their moral responsibility for overseeing the system's operation, misuse of automation features was common. For example, several drivers reported circumventing safety protocols by applying weight to the steering wheel to simulate hand presence. This behaviour reveals a deeper structural issue: moral responsibility alone is insufficient to guarantee adherence to safety protocols. Current system designs, which rely predominantly on basic compliance checks such as steering torque verification, inadvertently facilitate complacency and improper use.

To address this challenge, it is not enough simply to remind drivers of their responsibilities. Instead, automated systems must be proactively designed to promote continuous driver engagement and situational awareness. This includes implementing more robust human–machine interaction strategies that help ensure drivers remain alert and ready to assume control when necessary.

6.5.4 Limitations

Despite the insights gained from our research, several limitations may impact the interpretation and generalisability of our findings. First, the data used primarily reflect drivers' subjective perceptions. While these perceptions are valuable for understanding user experiences, they may

not always correspond to actual driving conditions. To address this limitation, we recommend that future research complement subjective reports with objective data (e.g., telemetry or kinematic data), allowing for statistical analysis that can more robustly assess system performance and user interaction.

Second, although all participants used Tesla's Autopilot or FSD Beta—both classified as SAE Level 2 systems—the original data collection did not ask participants to identify the automation level of their vehicle. While this could be seen as a limitation, we argue that it does not compromise the validity of our findings for three reasons: (1) users typically interpret automation through system behaviour and driver responsibility, rather than formal SAE terms; (2) knowledge-related questions in the interviews revealed that participants generally understood system limitations and the need for supervision (Nordhoff & Hagenzieker, 2024); and (3) Tesla communicates these limitations clearly through system prompts and manuals. Nonetheless, we recommend that future studies explicitly examine users' awareness of automation classifications or mental models, particularly where this may influence trust, expectations, or driver behaviour.

Third, our study is constrained by the fact that we only considered drivers and vehicle manufacturers as human agents in the evaluation of meaningful human control (MHC). However, it is important to acknowledge that other stakeholders—such as other road users, members of the public, lawmakers, and government authorities—also play a significant role in the operation, deployment, and governance of automated driving systems (Calvert & Mecacci, 2020). These stakeholders contribute to the broader sociotechnical context in which automated systems function. We therefore recommend that future research broaden the scope of analysis by including additional human agents to provide a more holistic evaluation of MHC.

Fourth, our evaluation of human expectations was primarily limited to safety. While safety remains a central concern in the context of driving automation, other expectations—such as comfort, regulatory compliance, and time efficiency—are also relevant. Future studies should incorporate a wider range of human expectations to enable a more comprehensive understanding of how MHC is established and maintained in partially automated systems.

Fifth, our findings are based exclusively on data collected from users of Tesla's FSD Beta programme in the United States and Canada. Variations in the design, implementation, and user interfaces of automated driving systems across manufacturers may lead to different user experiences and perceptions. Consequently, we recommend that future research include a more diverse sample of both automakers and participants to enhance the representativeness and generalisability of findings related to meaningful human control.

Sixth, while the dataset used in this study offers valuable insights into driver interactions with Tesla's Autopilot and FSD Beta systems, several potential biases should be acknowledged. Selection bias may have influenced the sample, as participants were recruited exclusively through online platforms. This recruitment method may have excluded individuals who are not active on such platforms, potentially leading to the underrepresentation of certain demographic groups. As a result, the sample may not fully reflect the diversity of all users of these systems. In addition, response bias may have affected the quality of interview data. Given the remote nature of the interviews (e.g., conducted via Zoom), participants may have tailored their responses to align with perceived social norms or desirability. These potential biases should be considered when interpreting the findings of this study.

Seventh, as a qualitative study, the analysis is subject to potential researcher bias. Although

the study employed structured analytical frameworks—such as inductive category development and the application of ad-SOTIF and PST assessments—the interpretation of participant responses necessarily involved a degree of researcher judgement. To enhance objectivity, future studies could incorporate inter-coder validation or triangulation methods to strengthen the reliability and transparency of qualitative coding processes.

Eighth, the retrospective and indirect nature of our evaluation of MHC poses an inherent methodological limitation. The dataset was initially collected without explicitly introducing the MHC framework or assessing participants' understanding of its core components—namely, tracking and tracing. As a result, we were unable to directly evaluate participants' awareness, interpretation, or valuation of MHC as a concept. While our indirect approach yielded contextually relevant and theoretically grounded insights, it was not originally designed with the explicit goal of evaluating MHC. Future research should therefore be conducted with the explicit intent to assess MHC—meaning that while similar questions might be asked, they would be purposefully framed within the MHC framework. This would allow for more focused interpretation, targeted measurement, and potentially more valid conclusions about how users understand and experience meaningful human control in automated driving contexts.

6.6 Conclusion

Evaluating meaningful human control (MHC) over partially automated driving systems presents considerable challenges, stemming from the complex interactions between human drivers and automation, as well as the variability inherent in real-world driving contexts. This study offers a systematic assessment of how such systems adhere to MHC principles beyond post-incident analysis and hypothetical scenarios, by focusing on their operation in everyday, real-world situations.

The contributions of this study are twofold. First, we evaluated the extent to which partially automated driving systems in real-world contexts comply with the requirements of MHC. Second, we introduced a novel methodological approach for assessing the tracking and tracing dimensions of MHC, using qualitative data derived from in-depth interviews with users of Tesla's Autopilot and FSD Beta systems. This approach provides richer insights into drivers' lived experiences with automation and offers a practical framework for examining MHC in automated systems.

We evaluated tracking based on how consistently various safety features performed their intended functions, alongside drivers' perceptions of safety and trust. Tracing was assessed through drivers' knowledge of the requirement to keep both hands on the steering wheel and remain alert, their capacity to execute corrective actions, and their awareness of moral responsibility for the system's operation. By applying this evaluation framework, we found that while subsystems of Tesla's FSD Beta and Autopilot demonstrate partial adherence to the principles of meaningful human control, several significant challenges remain. These include inconsistencies in tracking both driver and manufacturer safety expectations, as evidenced by the comparatively weaker tracking performance of F/SCW and AEB relative to BSM and LDA. Such issues are frequently linked to technological limitations—such as false positives, false negatives, and sensor vulnerabilities under adverse environmental conditions—as well as misaligned user expectations.

Inconsistencies in MHC also arise from variability in driver interaction, including selective adherence to safety protocols, over-reliance on automation, and misuse of system features. For example, adherence to guidelines—such as maintaining hand contact with the steering wheel and staying alert—was found to be inconsistent and often shaped by perceived risk. Drivers exhibited greater caution with FSD Beta in urban environments compared to the more relaxed use of Autopilot on highways.

These findings highlight the urgent need for further technological development, user-centred design improvements, and regulatory attention to ensure stronger and more consistent meaningful human control in partially automated driving systems.

Chapter 7

Objective Assessment of Meaningful Human Control

This chapter investigates how objective behavioural telemetry metrics relate to drivers' subjective experience of meaningful human control (MHC) in a simulated safety-critical scenario. In a driving simulator study, 24 participants interacted with two control modes, traded control (TC) and haptic shared control (HSC), and completed questionnaires assessing perceived meaningful human control while their driving behaviour was recorded. Overall, HSC yielded higher perceived control, understanding, cooperation, and safety compared to TC. Although only some of the hypothesised behavioural metrics reached statistical significance, several non-significant metrics were nonetheless supported by participants' qualitative explanations, indicating that these behaviours remained meaningful for how they formed their subjective evaluations.

Unlike Chapters 5 and 6, which retrospectively applied MHC to semi-structured interview data from real-world users of partially automated driving systems, this chapter uses a questionnaire specifically designed to translate the four actionable MHC properties derived from the tracking and tracing conditions (Cavalcante Siebert et al., 2023) into measurable evaluation criteria. These properties specify what a human-AI system must exhibit to be considered under meaningful human control. Building on them, this chapter derives both subjective perception items and objective behavioural telemetry metrics, allowing MHC to be evaluated through the relationship between what drivers experience and what their behaviour reveals. The simulator scenario reproduces the safety-critical conditions under which driver complacency might arise.

This chapter is based on the following paper (in preparation):
Suryana, L. E., George, A.*, Flipse, L., van Arem, B., Abbink, D., Calvert, S., Siebert, L. C., and Zgonnikov, The Illusion of Control? Linking Behaviour and Perception to Evaluate Meaningful Human Control over Partially Automated Driving.*

*Joint first authorship. L.E. Suryana led: Formal analysis (qualitative coding and inter-coder reliability assessment) and Methodology (design of initial interview questions and coding scheme). A. George led: Formal analysis (quantitative and statistical analysis) and Methodology (objective metrics and initial hypothesis formulation). Both authors contributed equally to: Conceptualization, Investigation, Data curation, Software, Validation, Visualization, writing – original draft, and Writing – review & editing.

7.1 Abstract

In this study, we investigated the extent to which drivers experience meaningful human control when interacting with automated driving systems. Twenty-four drivers completed a simulator study involving silent automation failures under two modes: traded control (TC) and haptic shared control (HSC). Participants rated their perceived control, and driving telemetry was converted into behavioural metrics. Subjective evaluations were collected after each scenario. Following the analysis plan, regression models tested the hypothesis that certain behavioral indicators would be related to perceived meaningful human control. While most of the hypothesised metrics did not show statistically significant relationships, qualitative analyses revealed further insights not captured by the quantitative data. For example, mismatched intentions and resistance to driver inputs, such as steering forces against the driver's preferred action, were commonly described as factors that undermined perceived control and the driver's sense of being understood by the system. The discussion synthesises both the quantitative and qualitative findings and provides guidelines for future research on evaluating meaningful human control in systems that involve physical interaction between humans and automated systems.

7.2 Introduction

Responsibility is generally considered to go hand-in-hand with the level of control (Flemisch et al., 2012). For example, the driver of a car has a much higher degree of responsibility and control as compared to a passenger in a train with a low degree of responsibility and control. This balance of responsibility and control can be disrupted when automation systems take up part of the control while the drivers of such vehicles are still held responsible for accidents (Beckers et al., 2022). Therefore, understanding how drivers perceive responsibility and control in partially automated vehicles is essential for the design of vehicle automation.

Vehicle automation is widely seen as a promising way to improve road safety and traffic efficiency by reducing human error and optimising vehicle behaviour (Fagnant & Kockelman, 2015). Early deployments of fully automated systems have demonstrated benefits, such as reductions in specific crash types (Kusano et al., 2025). In practice, however, most vehicles today operate at SAE Levels 2–3 (SAE International, 2021), where automation handles some driving tasks while human drivers must still supervise and intervene. These levels create a paradox: drivers are expected to remain constantly attentive even though their control role is significantly reduced (Endsley, 2017). This paradox is most evident in Advanced Driver Assistance Systems (ADAS), which support SAE Level 2 functions such as adaptive cruise control and lane centering (Dang et al., 2015; Wang et al., 2015).

One manifestation of this paradox is complacency, an over-reliance on automation in which drivers' reactions slow when manual intervention is required due to reduced vigilance. This decline occurs when automation shifts drivers from active control to passive supervision, reducing engagement (Chu & Liu, 2023). In human factors terms, this reflects a vigilance decrement caused by cognitive underload during passive monitoring (Endsley, 2017; Merat et al., 2014). Such disengagement is critical in scenarios where the driver must retake control unexpectedly: if unprepared, the driver may respond too slowly, increasing accident risk. At the same time, many commercial systems are promoted as being capable of "autonomous driving," even though

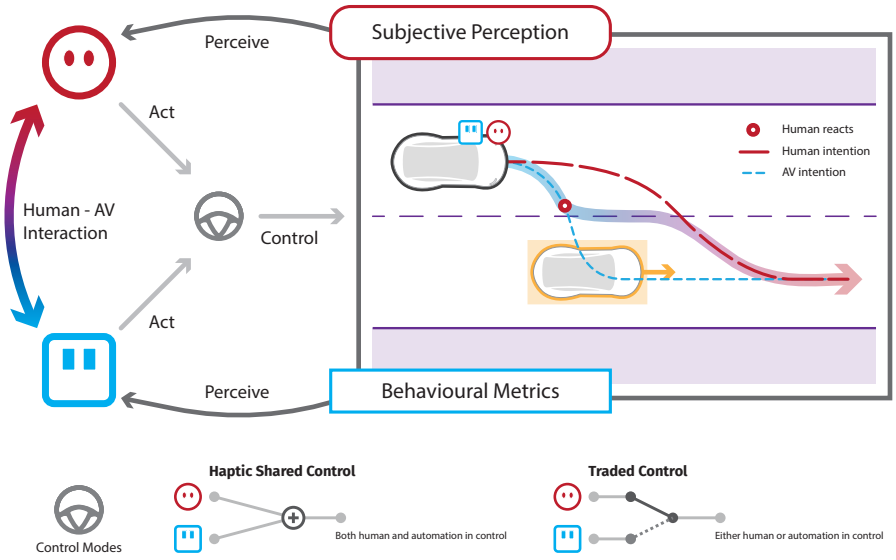


Figure 7.1: Understanding the relationship between the subjective perception of human drivers and behavioural metrics observable to machines is vital for designing automated systems under meaningful human control. This study explores how different control modes affect perceived control when human and automated systems have conflicting intentions.

manufacturers specify that drivers must remain ready to intervene when needed. This creates paradox of automation : the systems are marketed as relieving the driver of control, yet they simultaneously reduce the driver's ability to remain vigilant, raising questions about whether drivers can reasonably be assigned full legal responsibility under these conditions.

Beyond the paradox of automation, partial automation can also erode drivers' felt responsibility for vehicle behaviour, even though they remain legally accountable. This reduced sense of responsibility is often explained through the Sense of Agency (SoA), the subjective experience of initiating and controlling an action. Prior work shows that SoA tends to decline as automation increases, with higher autonomy reducing awareness and engagement (Moore, 2016; Cornelio et al., 2022; Berberian et al., 2012a). When individuals feel they did not cause an action, they are also less likely to accept moral responsibility for its outcomes (Moretto et al., 2011).

Taken together, complacency and declining SoA create a layered challenge: systems expect drivers to remain attentive and accountable, yet their design undermines the very capacities required to do so. This mismatch produces misattribution of responsibility (Matthias, 2004; Santoni de Sio & Mecacci, 2021), where it is unclear who should be held accountable for the actions of an automated system. Recent evaluations of commercial driving systems illustrate these gaps: drivers often lack understanding of how well the system can react—or how well they themselves can react—while manufacturers frequently distance themselves from responsibility through disclaimers (Calvert et al., 2020b).

To mitigate these responsibility gaps, meaningful human control (MHC) has been proposed as a normative stance that humans should have control and responsibility over the behaviour of automated systems(Santoni de Sio & Van den Hoven, 2018). Automated systems must meet two conditions to be under MHC — 1) the *tracking condition* which asserts that the behaviour of automated systems must track the reasons of relevant humans and 2) the *tracing condition*

which asserts that it should be possible to trace the responsibility for the behaviours of the automated system to at least one human. Based on tracking and tracing, subsequent deliberations have identified four actionable properties for a system to be under MHC: P.1) a clear *moral operational design domain*, P.2) *shared representations* between human and automation, P.3) humans have necessary *ability and authority to control* the behaviour of the automation, and P.4) that automation *actions are linked to a responsible human*(Cavalcante Siebert et al., 2023). Thus, as system being under MHC, provides grounds for assigning responsibility to different human actors.

While MHC has often been discussed as a guiding principle for system design Santoni de Sio & Van den Hoven (2018); Mecacci & Santoni de Sio (2020); Cavalcante Siebert et al. (2023); Calvert (2025), a necessary first step is to evaluate whether existing control strategies satisfy its conditions in practice. Without such evaluation, MHC risks remaining conceptually strong but practically under-validated. Previous studies have proposed several ways to evaluate MHC, but many of them assess it only indirectly. For example, some studies rely on perceptions derived from post-crash analyses (Calvert et al., 2020b) or from driving reports (Suryana et al., 2025b). However, evaluations based on post-crash analyses do not capture the nuances of real-time interaction between the driver and the automated system, while those based on driving reports rely on questions not specifically designed to measure MHC, even though partial alignment with MHC principles can be inferred.

In this work, we aim to capture drivers' perceptions of meaningful human control when interacting with automated driving systems. To achieve this, we combine a questionnaire explicitly designed to evaluate MHC with driving interaction data, following the approach by (Verhagen et al., 2024), who integrated subjective measures and interaction data to assess one aspect of MHC in human-robot interaction for firefighting systems. They argue that while subjective measures offer valuable insights into human experiences, objective metrics are essential to verify whether the system genuinely supports MHC. Such objective data helps assess the system's effectiveness in real-time decision-making scenarios and ensures that humans can be held accountable for robot actions, thereby preventing responsibility gaps. The key difference in our study is that we evaluate the entire MHC framework within the context of automated driving systems, where objective metrics derived from behavioral telemetry data—such as driver performance, interaction timing, and decision-making accuracy—provide critical evidence of MHC, complementing subjective perceptions.

To evaluate whether an automated vehicle embodies aspects of meaningful human control, we focus on driving control strategies currently implemented in automated vehicles. Previous work by (Suryana et al., 2024, 2025b) shows that current automated vehicles, while not initially designed with MHC principles in mind, nonetheless exhibit alignment with these conditions. This suggests that MHC can, to some extent, already emerge from existing system architectures. Accordingly, we use two common driving control strategies, Haptic Shared Control (HSC) and Traded Control (TC), as representative approaches to explore how such systems align with MHC in practice. In HSC, both driver and automation act simultaneously through force feedback on the steering wheel, supporting smoother collaboration and reducing conflict Abbink et al. (2018); Wang et al. (2017); Li et al. (2018). In contrast, TC relies on explicit handovers, which can be simpler to implement but risk disorientation if poorly timed de Winter et al. (2023). Figure 7.1 illustrates these control modes.

Given these existing systems and their potential alignment with MHC, this paper addresses

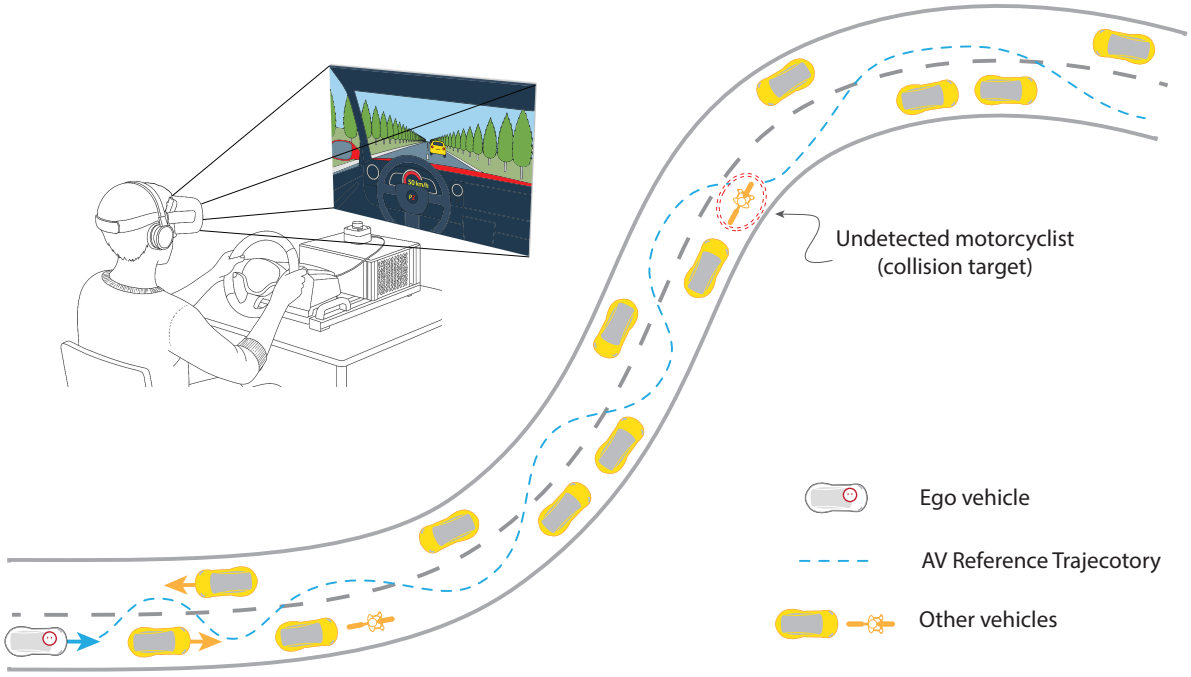


Figure 7.2: **Driving simulator experiment:** Experimental setup of the fixed-base driving simulator with VR headset, steering wheel, and audio system, and the repeated overtaking scenario simulating a silent automation failure on a two-lane, two-way road. The ego vehicle (white car) overtakes other vehicles and motorcyclists with oncoming traffic (yellow). To simulate silent automation failure, the automation would "fail to detect" and try to collide with one random motorcyclist per trial. The figure illustrates one representative configuration

the following research question: *To what extent do current driving control strategies enable drivers to experience meaningful human control in safety-critical situations?* We investigate this question in a controlled driving simulator study where participants interact with an automated vehicle under both HSC and TC conditions.

The main contributions of this paper are:

- An evaluation framework that integrates behavioural telemetry, subjective questionnaires, and qualitative insights to evaluate how drivers perceive Meaningful Human Control (MHC) when interacting with automated vehicles operating under different driving control strategies.
- Evaluating two control modes based on the driver's perception of control and responsibility.

7.3 Methods

We conducted a driving simulator experiment (Section 7.3.1) and used surveys to gauge drivers' subjective perception of interacting with driving automation through different control modes

(Section 7.3.2). These subjective measurements were complemented with objective behavioural metrics derived from the telemetry data (Section 7.3.3), which were further analysed to study the influence of control modes, and relations between subjective perceptions and behavioural metrics (Section 7.3.4).

Twenty-four participants (13 male, 11 female) between the age of 23 to 36 years (with average and SD 29.2 $\pm$ 3.75 years) were recruited from the student and research community at Delft University of Technology between December 2023 and April 2024 via flyers distributed through personal contacts and snowball sampling, where enrolled participants referred others. Eligible candidates must (i) have held a valid driving licence for at least one year, (ii) have normal or corrected-to-normal vision without spectacles (contact lenses were permitted), and (iii) have no history of epilepsy or other conditions that could be aggravated by virtual reality (VR).

The study protocol was approved by the Human Research Ethics Committee (HREC) of Delft University of Technology (ID: 111053). Written informed consent was obtained from all participants prior to the experiment. At the beginning of the experiment, participants were reminded of their right to withdraw at any time without penalty. A €10 voucher was provided to each participant upon completion of the experiment. All data were anonymised and stored using unique participant identification codes.

7.3.1 Driving simulator experiment

The participants were asked to drive through a section of a rural road supported by driving automation; this was done repeatedly over a sequence of trials. In each trial, participants had to complete a sequence of overtaking manoeuvres while interacting with an automated driving system through a haptic steering wheel (SensoDrive high fidelity steering wheel SensoDrive (2025)). Participants were presented a first person-view using Varjo VR-3 virtual reality headset while seated in front of a steering wheel which was capable of providing haptic feedback (Figure 7.2). The experiment was configured in JOAN Beckers et al. (2023) – the framework for running experiments in the CARLA environment Dosovitskiy et al. (2017). Sony WH-1000XM3 noise-cancelling headphones were used to reduce auditory distractions.

As the participants drove through a bidirectional rural road with straight and winding sections, the speed of the ego vehicle was fixed at 50 km/h under cruise control and the participants had no control over the accelerator or brakes. They could only control the steering wheel. While driving they encountered right-hand traffic travelling in both directions at a constant speed of 40 km/h, nine travelling in the same direction, and five in the opposite direction (Figure 7.2). The presence of oncoming traffic ensured that drivers had to continuously switch lanes to avoid collisions. In each trial, participants were responsible for controlling the steering to prevent collisions with other vehicles.

Apart from trials with manual control, where the driver had complete control over the steering, there were trials where the driver had to interact with a driving automation system. The driving automation was programmed to follow a pre-recorded reference trajectory. Participants interacted with the automation in two control modes (varied across trials): **haptic shared control (HSC)** and **traded control (TC)**. In HSC, both the automation and the driver can apply torques on the steering wheel at the same time and thus steer the vehicle together. In TC, if

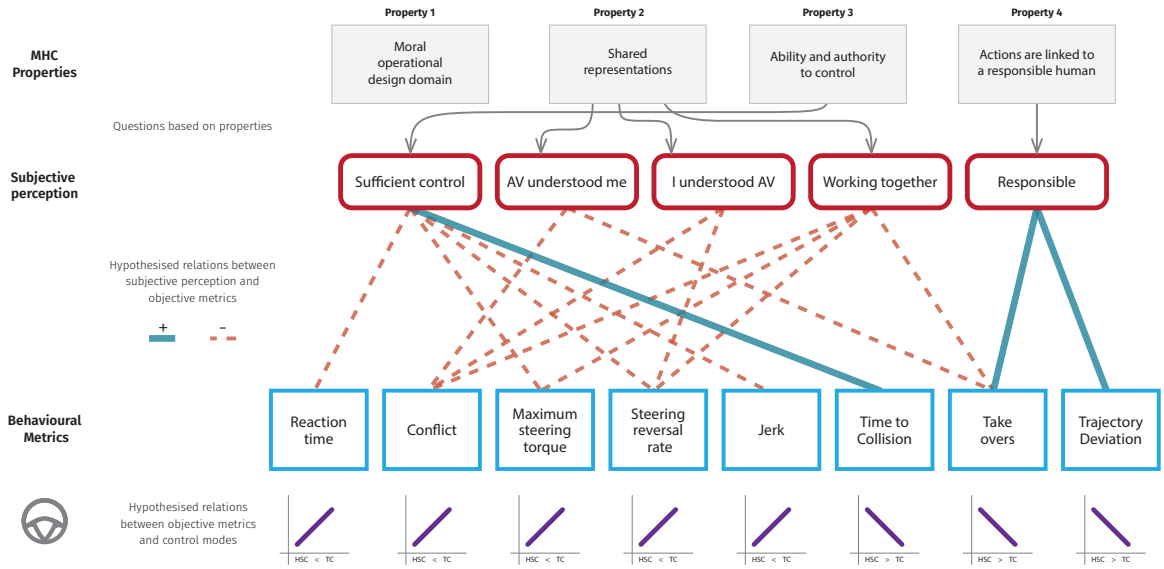


Figure 7.3: **Deriving hypotheses about survey questions and behavioural metrics from the properties of systems under meaningful human control (MHC)**: Survey questions related to MHC properties were used to evaluate subjective perception of MHC. The figure also shows the hypothesised dependence of subjective perception of MHC on behavioural metrics and control modes.

the driver applies a torque above a threshold, the automation turns off for one second while the driver has complete authority over steering the ego vehicle. If the torque applied by the human remains below the threshold for more than one second, automation torque gradually increases until it regains full authority or is intervened upon by the driver.

To familiarise participants with the control modes and driving task, each participant completed four familiarisation trials: (i) manual driving without traffic, (ii) manual driving with traffic, (iii) driving with traffic under HSC, and (iv) driving with traffic under TC. The main experiment then consisted of nine trials: Trial 1 involved manual driving, while Trials 2–9 featured automated driving in either TC or HSC, with four trials per mode. The order of automated trials was randomized to balance conditions across participants.

To investigate participants' interaction with automation in safety-critical scenarios, we simulated a silent automation failure in every automated driving trial. Specifically, the ego vehicle was programmed to follow a trajectory leading toward a potential side collision with a motorcycle as illustrated in Figure 7.2. The motorcycle's position was randomized across trials to avoid learning effects, with position distributions balanced across participants and control modes.

Participants were instructed to remain in their lane unless overtaking is needed, avoid collisions, stay on the road, and retain responsibility for safety because driving automation is imperfect and failures are possible. Participant's subjective perceptions were quantified after each trial using a post-trial survey (Section 7.3.2). After completing all the trials, a descriptive post-experiment questionnaire was used to qualitatively assess the subjective perceptions about the interaction (Section 7.3.2).

7.3.2 Subjective perception of MHC

Post-trial surveys with a rating scale were used to quantify the perception of the participants (Section 7.3.2) and open ended post-experiment surveys were used to get a deeper understanding of the cues that influenced to their perception (Section 7.3.2).

Post-trial subjective scores

Seven Likert-scale (1 to 10) questions (Table 7.1) were asked after each trial to gauge how participants experienced the interaction with the driving automation and to ultimately assess the extent to which the driving automation operated under MHC. The questions were inspired from the actionable properties of systems under MHC Cavalcante Siebert et al. (2023). Since we were more focussed on the interaction between the human and the automation, rather than on the design of the automation itself, we designed the control modes HSC and TC to have identical moral operational design domains (moral ODDs). Hence, we did not include any question for Property 1 (*moral operational design domain* (moral ODD)). Property 2 (*shared representations*) was divided into three components: **AV understood me**, **I understood AV**, and **working together**. This division reflects Cavalcante Siebert et al.'s Cavalcante Siebert et al. (2023) description that shared representations between human and AI systems depend on how both agents understand each other and are able to update their representations in response to changing reasons. The remaining properties were each represented by one subjective perception item: **sufficient control** for Property 3 (*ability and authority to control*) and **responsible** for Property 4 (*actions are linked to a responsible human*). Figure 7.3 shows the mapping between MHC properties, subjective perception questions, and behavioural metrics.

In addition to items directly related to the MHC properties Cavalcante Siebert et al. (2023), we included questions on perceived **safety** and **trust**; as previous research has demonstrated, they can strongly influence evaluations of MHC based on how drivers experience interactions with driving automation Suryana et al. (2024).

As the manual driving trials did not include driving automation, some of the questions were not applicable to these trials and a modified set to questions M1, M2 and M3, were used evaluate the baseline perception of sufficient control, responsibility and safety respectively (Table 7.1).

Post-experiment descriptions

After all the driving trials, an open-ended post-experiment questionnaire was administered to get a deeper understanding of the perceptions of the participant and the factors that might have influenced these perceptions. The questions related to the same concepts mentioned in post-trial questions (*sufficient control*, *AV understood me*, *I understood AV*, *working together responsibility*, *safety*, and *trust*). Two questions per concept probed whether they had a positive or negative experience of that concept and to detail their experience. For example, the questions related to *sufficient control* were:
D1: "Were there any situations where you felt you *had sufficient control* over the automated vehicle operation? Please describe." and,
D2: "Were there any situations where you felt you *did not have sufficient control* over the automated vehicle operation? Please describe."

Post-trial questions – automated driving trials	**Property**
A1: I felt that I had sufficient control over the automated vehicle	Sufficient control (P3)
A2: I felt that the automated vehicle understood my intentions during the driving task	AV understood me (P2)
A3: I felt that I had sufficient understanding about the behaviours of automated vehicle	I understood AV (P2)
A4: I felt that the automated vehicle and I were working together towards the same goal	Working together (P2)
A5: I felt responsible for the driving task when I was using the automated vehicle	Responsible (P4)
A6: I felt safe in the automated vehicle during the driving task	Safe
A7: I trusted the automated vehicle during the driving task	Trust
Post-trial questions – manual driving trials	
M1. I felt that I had sufficient control over the vehicle.	Sufficient control
M2. I felt responsible for the driving task when I was using the vehicle.	Responsible
M3. I felt safe in the vehicle during the driving task.	Safe

Table 7.1: Post-trial questions that participants had to answer on a Likert scale (1 to 10) after interacting with the automated driving system through haptic shared control or traded control, and their relation to the four properties for meaningful human control. The questions for manual driving were redacted since there was no automation involved.

An exhaustive list of the questions is included in the supplementary material.

These items were designed to record in greater detail the overall impressions of participants for HSC and TC, to elicit scenarios that positively or negatively affected their perception of the qualities being measured in the post-trial questionnaire, and to compare their experiences across conditions. In particular, they addressed aspects that could not be evaluated meaningfully on a trial-by-trial basis, such as general system preference, system characteristics that affect workload, and qualitative reflections on system behaviour. This approach follows Verhagen et al. (2024), who employed post-experiment reflections to evaluate the traceability dimension of MHC. In our study, such qualitative insights provide valuable context for understanding the Likert-scale scores recorded after each trial.

7.3.3 Behavioural metrics and hypotheses

To study how these subjective perceptions were related to behavioural aspects of driver-automation interaction, we formulated behavioural metrics which were hypothesised to be correlated to subjective perception scores 7.3. These metrics quantify aspects related to the interaction dynamics between the driver and the automation, driver's performance, and the characteristics of vehicle trajectories. Brief descriptions of metrics and the hypotheses linked to each are mentioned below.

Reaction time: In experiment trials, where a silent automation failure would trigger a risky manoeuvre towards a motorcyclist, the reaction time was quantified as the time between the initiation of the risky manoeuvre and first instant when the steering torque applied by the driver exceeded a threshold. We hypothesised that smaller reaction times would be correlated with greater scores for 'sufficient control', and that the reaction times would be shorter for HSC than

for TC.

Conflict in steering torques: In the experiment, the human driver interacted with the automation through the steering wheel and we use the conflict in steering torques to measure the disagreement between the human and the automation. In line with prior research (Boink et al., 2014), we hypothesise that high values of conflict are negatively correlated with scores for 'AV understood me'. 'I understood AV' and 'Working together'. Regarding the control modes, based on the assumption that interactions with HSC will be smoother, we hypothesis that conflict will be lower for HSC than for TC.

Maximum steering torque: When interacting with the automation, the maximum steering torque exerted by the participant reflect their control effort. Earlier studies have also shown that drivers exert maximum torque when resisting lane-keeping or lane-departure assistance (Ercan et al., 2018). Thus, we hypothesise that higher maximum steering torques are negatively correlated with scores for 'sufficient control' and 'working together'. Furthermore, owing to the smoothness of the interaction in HSC, we hypothesise that maximum steering torques will be lower for HSC than in TC.

Steering reversal rate: A established method for quantifying the control effort of a driver considers the steering reversal rate: the greater the steering reversal rate, the greater the control effort of the driver (Mars et al., 2014). We hypothesise that higher steering reversal rates are negatively correlated with the scores for 'sufficient control', 'I understood AV' and 'Working together'. Based on the continuous nature of HSC, we hypothesise that the steering reversal rate would be lower for HSC than for TC.

Jerk of vehicle trajectories: The smoothness of the trajectories of the ego vehicle were also analysed based on the root mean square jerk of the ego trajectory. We hypothesised that non-smooth trajectories with higher jerks would be negatively correlated with the score for 'sufficient control'. Also, assuming that HSC would result in smoother trajectories, we hypothesised that jerk would be smaller for HSC than TC.

Time to Collision (TTC): In each trial where a simulated automation failure could lead to a potential collision with a motorcyclist, the minimum TTC betwween the ego vehicle and the motorcyclist can be used to measure the criticality of an interaction: the lower the TTC, the more critical the interaction is. We hypothesised that greater criticality of interactions (with low TTCs) would negatively impact the sense of 'sufficient control' i.e., that high values of TTC would be correlated with high scores for 'sufficient control'. Furthermore, based on the assumption that drivers can react sooner in HSC leading to less critical situations, we hypothesise that TTC would be larger for HSC than for TC.

Number of takeovers: When a driver is interacting with the automation, a take over is said to have happened if the torque applied by the driver exceeds the threshold which would disengage the automation in TC. We assumed that takeovers would be triggered when the behaviour of the automation were not aligned with the intentions of the driver and hypothesised that the number of take overs would be negatively correlated with scores of 'AV understood me' and 'working together'. At the same time, we hypothesised that the number of takeovers would be positively correlated with the scores for 'responsible', as the driver would have more influence on the trajectory of the ego vehicle with more take overs. We assumed that participants would be more engaged in HSC and hypothesised that the number of takeovers would be greater in the case of HSC than for TC.

Trajectory deviation: The deviation of the trajectory of the ego vehicle from the reference trajectory directly reflect the contribution of the driver. Thus, we hypothesise that trajectory deviation is positively correlated with the score for 'responsible'. Also, assuming that drivers are more engaged in HSC, we hypothesise that trajectory deviation is higher for HSC than TC.

Overtaking time: In our exploratory analysis, to quantify how driver's preferences might deviate from the reference trajectory, we also included the overtaking time defined as the time spent by the ego vehicle in the overtaking lane. Since, the preferences of individual drivers might lead to longer or shorter overtaking times, we did not associate any hypotheses with the overtaking time.

All of these metrics by providing objective insight into the behaviour of drivers, complement the data collected about the subjective perception of various concepts related to MHC. More detailed descriptions about the definition and calculation of these metrics can be found in the supplementary materials [1].

7.3.4 Analysis

Behavioural data and subjective scores collected during the experiment was analysed quantitatively to test hypotheses and the answers to post-experiment questionnaires were qualitatively analysed to garner deeper insights into factors affecting them.

Quantitative Analysis

To analyse the effect of the control modes (HSC and TC) on the behaviour of participants, we fit linear mixed-effect models (LMMs) for each behavioural metric (Section 7.3.3) with the z-scored behavioural metric as the dependent variable, control mode as the independent variable with a fixed effect, and participant ID as a random intercept.

To examine how participant's perceptions during the experiment were related to their behaviour and control modes, we analysed the subjective scores which were recorded in the post-trial questionnaire (Section 7.3.2). For each subjective perception item, one LMM was fit with the subjective score as the dependent variable, control modes and all behavioural metrics as independent variables with fixed effects, and participant ID as a random intercept.

The results of these LMMs fit for behavioural metrics and subjective scores were used in the confirmatory analyses of the hypotheses (Figure 7.3). Since two sets of models were fit on the data, one for each behavioural metric and subjective score, we used a Bonferonni correction of 2 to arrive at $p < 0.025$ $(= 0.05/2)$ for testing hypotheses for the confirmatory analysis.

Besides the confirmatory analysis, we conducted an exploratory analysis of the LMMs of the subjective scores to identify any relations between subjective scores and behavioural metrics that might have been overlooked in the hypotheses. The purpose of the exploratory analysis was not to make further claims, but to provide fodder for future experiments. In our initial hypothesis, we had assumed that the relationships between subjective scores and behavioural metrics would be the same across control modes. But some responses from in the post-experiment ques-

[1] Link to supplementary materials: under preparation

tionnaire hinted that the nature of the interaction, as determined by the control mode, might also influence the relationships between subjective scores and behavioural metrics. To study whether control modes influence these fixed-effects, we split the data and fit separate LMMs for each control mode. For the exploratory analysis, statistical significance of slopes was assessed at $p < 0.05$.

Qualitative analysis

Free-text answers of participants to post-experiment questionnaires were analysed by two coders. First, both coders reviewed participants' responses in their entirety to ensure familiarity with the dataset. An initial coding pass was conducted by one coder to identify all possible topics mentioned by participants, without restricting the scope to pre-defined categories. This exploratory, data-driven approach allowed for the inclusion of topics beyond those anticipated by the MHC framework.

Following the initial coding, similar topics were consolidated into broader thematic categories. Within each category, related ideas were organised into subtopics that together captured the range of participant experiences. This hierarchical structure enabled the preservation of nuanced perspectives while reducing redundancy and complexity in the dataset.

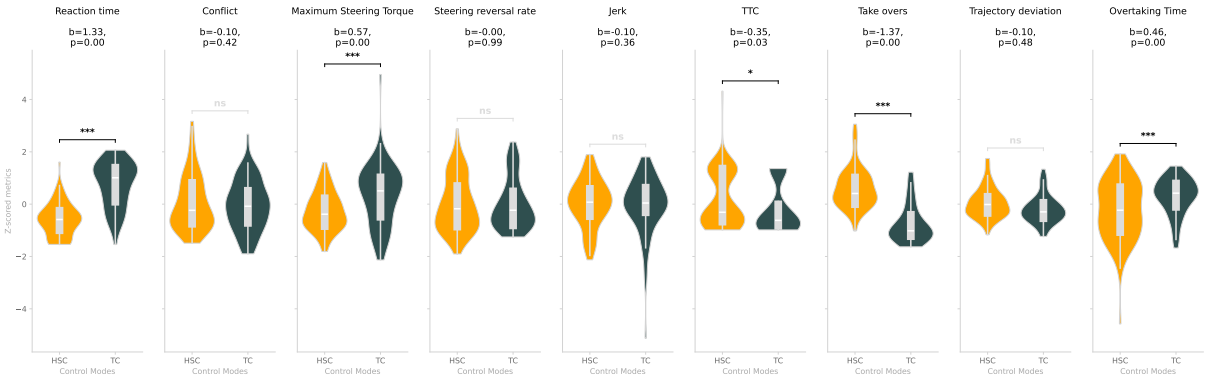
Once the preliminary categorisation was complete, a second coder independently reviewed the grouped topics. Both coders then developed labels and classifications for each factor, determining the most appropriate and precise naming. This independent labelling stage ensured that classification decisions were made without bias from prior discussion.

Inter-coder agreement was assessed on a set of 53 coded items. The observed agreement between coders was 79.2%, with Cohen's kappa (Cohen, 1960) of –0.066 and Gwet's AC1 (Gwet, 2008) of 0.745. The negative kappa value was attributable to the absence of "No" cases in the ground truth, which can produce prevalence bias in kappa calculations (Feinstein & Cicchetti, 1990). The agreement between both coders was on the presence of a factor in 42 cases, with disagreement in 11 cases (nine coded as present by Coder 1 only, two coded as present by Coder 2 only), and no instances where both coded the factor as absent. Given these conditions, Gwet's AC1 was taken as the more robust measure of agreement.

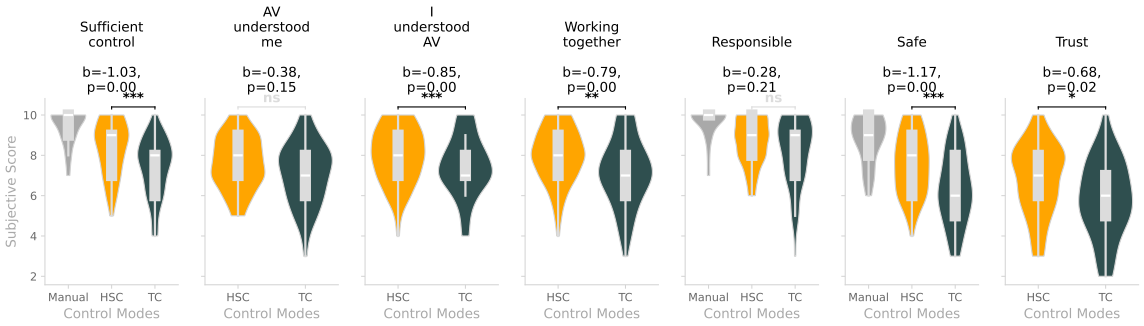
Discrepancies between coders were resolved through discussion until full consensus was achieved. This consensus-based, investigator-triangulation approach follows established best practices for directed content analysis (Hill et al., 2005; Braun & Clarke, 2006; Campbell et al., 2013).

7.4 Results

We present the quantitative results from the perspective of the behavioural metrics and their relation to the control modes and subjective scores (7.4.1) and the qualitative results from the perspective of subjective perceptions (7.4.2).



(a) **Behavioural metrics vs control mode.** Estimated differences in behavioural metrics between HSC and TC, controlling for participant-level baselines is represented by the slope coefficients for LMMs fit with the formula: behavioural-metric $\sim$ control-mode $+ (1|\text{participant})$.



(b) **Subjective perception vs control modes**. Estimated differences in perception scores between HSC and TC, controlling for participant-level baselines is represented by the slope coefficients for the LMM fit with the formula: subjective-score $\sim$ behavioural-metric $+$ control-mode $+ (1|\text{participant})$.

Figure 7.4: **Quantitative results for control modes.** This figure summarises how control mode influenced behavioural metrics (a) and subjective scores (b). Only the results in (a) were used to test the hypotheses about the effect of control modes on behavioural metrics (Table 7.2) at $\alpha = 0.05/8 = 0.006$. In this figure, * indicates $p < 0.05$.

7.4.1 Quantitative findings

The statistics of behavioural metrics and subjective scores with respect to the control modes are presented in Fig. 7.4. The results of the confirmatory analysis regarding the hypothesised relationship behavioural metrics and control modes are shown in Table 7.2. Summary of the confirmatory analysis of the hypothesised correlations between behavioural metrics and subjective score are shown in Table 7.3. The results from the exploratory analysis for uncovering correlations between subjective scores and behavioural metrics are collected in Fig. 7.5. We highlight the main findings for each behavioural metric below.

Reaction time: In accordance with our hypothesis, reaction times were significantly lower for HSC than for TC ($\beta = 1.33$, $p < 0.001$). However, contrary to our hypothesis, reaction time was positively correlated with the score for 'sufficient control' ($\beta = 0.24$, $p = 0.01$). The exploratory analysis also revealed a positive correlation between reaction time and the score of 'responsible' for HSC ($\beta = 0.18$, $p = 0.04$).

Behavioural Metric	Hypothesised relations	Observed relations	Slope β	p-value	Hypothesis accepted
Reaction time	HSC < TC	<	1.33*	< 0.001	✓
Conflict	HSC < TC		−0.10	0.42	
Maximum steering torque	HSC < TC	<	0.57*	< 0.001	✓
Steering reversal rate	HSC < TC		−0.00	0.99	
Jerk	HSC < TC		−0.10	0.36	
TTC	HSC < TC		−0.35	0.03	
Take overs	HSC < TC	>	−1.37*	< 0.001	✓
Trajectory deviation	HSC < TC		−0.10	0.48	

Table 7.2: **Confirmatory analysis – behavioural metrics vs control modes:** Hypothesised relations between behavioural metrics and control modes, which were tested at $p < 0.025$, from the results of the LMM fits for behavioural-metric $\sim$ control-mode $+ (1|\text{participant})$. '>' indicates greater values of the metric for HSC than for TC and '<' indicates the opposite. To control for family-wise error error rate when testing the eight hypotheses a Bonferroni correction was applied on $\alpha = 0.05/8 = 0.006$ as the level of significance.

Conflict in steering torques: Even though the conflict in steering torques for HSC and TC were not significantly different, we did find evidence for our hypothesis that conflict is negatively correlated with the score for 'AV understood me' ($\beta = -0.34$, $p = 0.001$). Furthermore, the exploratory analysis showed that conflict was positively correlated with the score of 'sufficient control' for HSC ($\beta = 0.39$, $p = 0.01$) and negatively correlated with the score of 'responsible' for TC ($\beta = -0.27$, $p = 0.03$).

Maximum steering torque: Consistent with our hypothesis, maximum steering torque was lower for HSC than for TC ($\beta = 0.57$, $p < 0.001$). The exploratory analysis also showed a positive correlation between maximum steering torque and the score of 'responsible' for TC.

Steering reversal rate: For steering reversal rate no significant differences between control modes or correlations with subjective scores were found.

Jerk of vehicle trajectories: The data showed no significant difference between the jerk of vehicle trajectories between HSC and TC. None of the hypothesised correlations of jerk with subjective scores were substantiated in the confirmatory analysis. The exploratory analysis showed that jerk was negatively correlated with the score of 'safe' ($\beta = -0.22$, $p = 0.01$) — especially for TC ($\beta = -0.314$, $p = 0.01$).

Time to collision: The data did not show a significant difference in TTC for the two control modes HSC and TC. There was no evidence the hypothesised correlation between TTC and the score of 'sufficient control' either. However, the exploratory anaylysis did show a positive correlation between TTC and the score for 'AV understood me' for TC.

Number of takeovers: As per our hypothesis, the number of takeovers was larger for HSC than for TC ($\beta = -1.37$, $p < 0.001$). No evidence was however seen for correlations of the number of takeovers with any of the subjective scores.

Trajectory deviation: Trajectory deviation showed no significant difference between HSC and TC. The confirmatory analysis could not find evidence for any of the hypothesised correlations

Subjective score	Behavioural metric	Hypotheses	Slope β	p-value	α	Hypothesis accepted
Sufficient control	Reaction time	-	*0.238**	*0.014*	0.001	
	Maximum steering torque	-	0.038	0.666	0.001	
	Steering reversal rate	-	-0.178	0.086	0.001	
	Jerk	-	-0.148	0.083	0.001	
	TTC	+	0.033	0.621	0.001	
AV understood me	Conflict	-	*-0.344**	*0.001*	0.003	✓
	Take overs	-	-0.019	0.880	0.003	
I understood AV	Conflict	-	-0.130	0.203	0.003	
	Steering reversal rate	-	-0.038	0.720	0.003	
Working together	Conflict	-	-0.110	0.307	0.001	
	Maximum steering torque	-	-0.080	0.402	0.001	
	Steering reversal rate	-	0.061	0.565	0.001	
	Take overs	-	-0.176	0.167	0.001	
Responsible	Take overs	+	0.182	0.90	0.003	
	Trajectory deviation	+	-0.022	0.769	0.003	

Table 7.3: **Confirmatory analysis — subjective scores vs behavioural metrics:** Hypothesised relations between subjective answers and behavioural metrics which were tested at $p < 0.025$ for the LMM fits for subjective-score $\sim$ behavioural-metric + control_mode + (1|participant). '+' indicates that perception score increases with an increase in the metric and '-' indicates a reduction in perception score decreases with an increase in the metric. For each subjective score, when testing the h_s hypotheses related to it, a Bonferroni correction was applied to control the family-wise error rate, with $\alpha = 0.05/h_s$ as the level of significance.

of trajectory deviation with subjective scores. The exploratory analysis showed that trajectory deviation was negatively correlated with the score of 'sufficient control' ($\beta = -0.19$, $p = 0.02$) and positively correlated with the score for 'trust' ($\beta = 0.24$, $p = 0.01$) — especially for TC ($\beta = 0.29$, $p = 0.05$).

Overtaking time: Overtaking time was not part of our hypotheses and was purely included for exploratory purposes. Data showed that overtaking time was larger for TC than for HSC ($\beta = 0.46$, $p < 0.001$). Overtaking time was also found to be positively correlated with safety($\beta = 0.22$, $p = 0.04$) and trust ($\beta = 0.24$, $p = 0.05$). Additionally, for HSC, there was a postive correlation between overtaking time and the score of 'sufficient control' ($\beta = 0.35$, $p = 0.01$).

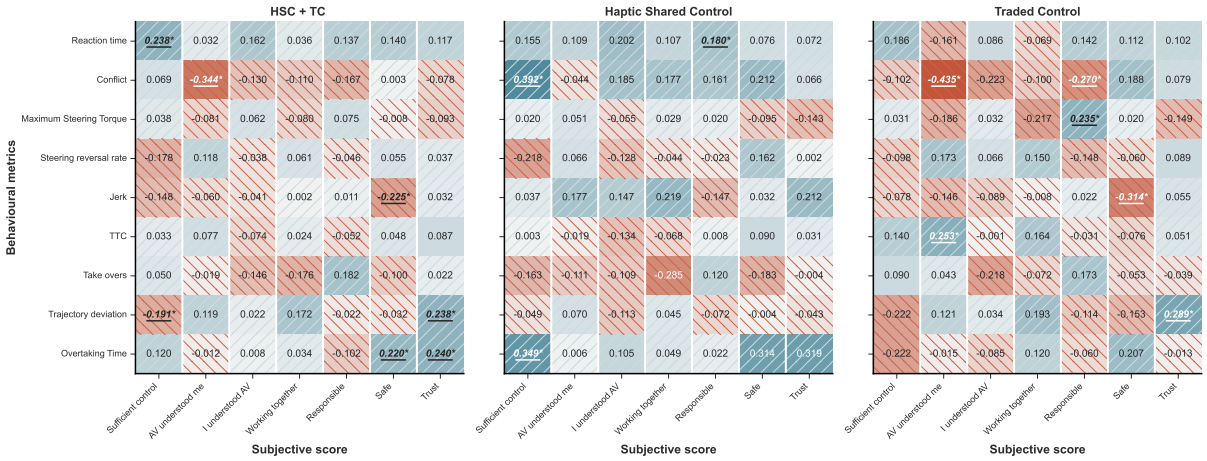


Figure 7.5: **Exploratory analysis — Subjective scores vs behavioural metrics:** a) The slope coefficients for the metric from the mixed-model regression with the formula: subjective-score $\sim$ behavioural-metric $+$ control-mode $+ (1|\text{participant})$. * indicates $p < 0.05$.

7.4.2 Qualitative findings

From the qualitative analysis, we identified factors that positively and negatively affected different perceptions related to MHC. These factors along with remarks for control modes summarised in Table 7.4. We classified the factors into three categories: **(1) epistemic factors** ● relating to mental states of participants, **(2) interaction factors** ■ relating to human-av interaction through the steering wheel, and **(3) trajectory factors** ▶ relating to the motion of vehicles. We follow this categorisation when describing the findings for each type of perception are described below.

Sufficient Control

Within the **epistemic** dimension, participants reported that HSC gave them a greater sense of control, describing that in most cases they felt they had sufficient control over the AV's operation (ID61, ID90). Regarding **trajectory**, one participant noted limitations in their ability to influence certain driving features, particularly vehicle speed, as a factor that reduced their perception of having sufficient control (ID58). In terms of **interaction**, the ease of overriding the AV's actions and maintaining continuous access to the steering wheel were identified as factors that enhanced the perception of control (ID17, ID58, ID172). Conversely, situations where the system applied steering forces against the driver's intention, in both TC and HSC (ID106, ID203), created a sense that control was being taken away rather than shared. Furthermore, the time and effort required to readapt to manual control after abrupt takeovers (ID61, ID90, ID172, ID242), the need for continuous vigilance (ID19), and the exertion of invasive forces in TC were cited as reducing the sense of control.

AV understood me

Perception of whether the AV understood the driver often depended on its ability to anticipate intentions accurately and act safely. In trajectory, participants reported that when the vehicle anticipated their manoeuvre intentions, such as overtaking or aligning within the lane, they felt the AV understood them (ID167, ID172). Conversely, mismatched anticipation, such as returning to the lane too early and creating a potential crash situation, made them feel misunderstood (ID229). From the perspective of human–automation interaction, low steering resistance (ID17) and the perception that TC would give control the moment the driver intervened (ID61) were taken as behaviours that aligned with the perception that the system understood the driver. On the other hand, situations where drivers felt the system applied strong steering resistance against their preferences (ID264) reduced this sense of understanding. Finally, in terms of epistemic factors, TC was sometimes linked to unsafe perceptions when they changed lanes while a motorcycle still adjacent (ID106, ID172).

I understood AV

Understanding of the AV's behaviour was often shaped by the predictability of its actions. In terms of epistemic factors, participants noted that familiarity with how HSC and TC operated made them understand the AV's intentions (ID17, ID58). For trajectory, several negative experiences were reported, including erratic actions such as briefly driving off the road (ID302), unsafe manoeuvres that could lead to a crash (ID454), and mismatched manoeuvre timings that were either too early or too late (ID137, ID203). These actions influenced participants' confidence in anticipating the AV's next move. Regarding human–automation interaction, subtle torque feedback was perceived as a helpful indicator for understanding the AV's intentions (ID58, ID61), while overly strong torque feedback (ID172) or unnecessary steering jerks (ID17) were seen as intrusive factors.

Working together

In epistemic factors, a sense of cooperation between the driver and the AV was associated with HSC (ID172). In trajectory, safe manoeuvres such as avoiding a bump with the vehicle in front (ID227) were viewed as signs of effective teamwork, whereas unsafe manoeuvres that could potentially lead to a crash with other vehicles undermined this perception (ID224). For human–automation interaction, subtle steering corrections provided by the AV, the ability to take over control with minimal pressure in HSC (ID17), and a sense of shared execution (ID40) reinforced the feeling that participants and the AV were working together. However, one participant also described having to make steering corrections themselves due to mismatched intent, which was seen as undermining the feeling of working together with the automation systems (ID58).

Table 7.4: **Factors affecting the subjective perception:** The qualitative analysis of post-experiment answers revealed factors that participants believed to *positively* and *negatively* affect their subjective perception. Remarks pertaining to the nature of the *control modes* were also summarised. These factors were classified into three categories as 1) those pertaining to *epistemic* ● states, 2) those pertaining to the *interaction* ■ between the human and the automation, and 3) those pertaining to the *trajectory* ▶ of the vehicle.

Positive	Negative	Control mode
Sufficient Control		
	● Need for continuous vigilance	● Greater sense of control for HSC
■ Continuous access to control (HSC)	■ Abruptness of take overs (TC)	
■ Override capability	■ Force against driver intention	■ Forces in TC more invasive
	■ Control overshoots	
	■ Disruptions in fluency	
AV understood me		
● Anticipation of intention	● Mismatched anticipation of intention	
■ Subtle torque feedback	● Lack of safety (TC)	
■ Low steering resistance	■ High steering resistance	
■ Override capability (HSC)		
▶ Timing	▶ Mismatched timing	
▶ Smooth lane alignment		
I understood AV		
● Transparency of AV intention	● Lack of transparency of AV intention	
■ Clear takeovers	● Lack of safety	
■ Subtle torque feedback	■ Stronger torque feedback	
	■ Unnecessary steering jerks	
	▶ Erratic trajectories	
	▶ Mismatched timing	
Working together		
● Effortless coordination (HSC)		● Drivers preferring autonomy, preferred HSC
● Shared direction (HSC)		
● Safety	● Lack of safety	
■ Subtle AV corrections matching driver intent	■ Driver corrections due to mismatched intent	

Table 7.4: **Factors affecting the subjective perception:** The qualitative analysis of post-experiment answers revealed factors that participants believed to *positively* and *negatively* affect their subjective perception. Remarks pertaining to the nature of the *control modes* were also summarised. These factors were classified into three categories as 1) those pertaining to *epistemic* ● states, 2) those pertaining to the *interaction* ■ between the human and the automation, and 3) those pertaining to the *trajectory* ▶ of the vehicle.

Positive	Negative	Control mode
Responsible		
● Need for supervision (TC), & Expectation to correct (HSC)	● Redundancy of human input	● Control mode defined responsibility
● Different driving style	● Shared-intent moments (TC)	
● Safety-critical situations	● Monotonous driving	
● Novelty and untrustworthiness		
● Accountability (moral, legal)	■ Physically disengaged (TC)	
■ Low steering resistance	■ Force against driver intention (HSC)	

Responsibility

Participants' awareness of responsibility was shaped by how they understood their role, the driving situation, and the nature of their interaction with the AV. In epistemic factors, some participants recognised their operational responsibility to supervise in TC and to drive as in manual mode when using HSC (ID61), both of which positively influenced their awareness of responsibility. Others acknowledged knowing they were the responsible party even when they were not actively controlling the vehicle (ID172). Counterintuitively, a low level of trust in the system positively influenced participants' perception of responsibility, with some attributing this to their first time driving such a system (ID203). However, a perception of partial responsibility, in which the participant could intervene with the AV but also recognised that the AV could complete the manoeuvre without their input, led to ambiguity that reduced their feeling of responsibility (ID17). In terms of trajectory, driving situations that were safety critical, such as overtaking (ID430), led to a higher perception of responsibility, while monotonous driving on a straight road (ID278) and low traffic situations (ID203) were associated with feeling less responsible. Finally, for human automation interaction, both HSC and TC were reported as causing negative perceptions of responsibility. In TC, participants felt that they could let the steering wheel operate by itself, which reduced their sense of responsibility (ID90). In HSC, one participant reported that feeling forced to make certain moves made them feel less responsible (ID17).

7.5 Discussion

We synthesised the information from qualitative post-experiment questionnaires, quantitative post-trial surveys and behavioural metrics to get a deeper understanding of the factors affecting the perception of control and responsibility of drivers interacting with an automation system in Section 7.5.1. We follow this with a discussion on the comparison of control modes in Section 7.5.2. Implications of the MHC framework and guidelines for applying it are discussed in Section 7.5.3. We wrap up with the limitations of this work and directions for future research in Section 7.5.4.

7.5.1 Factors related to subjective perception

The quantitative analysis only confirmed one of our hypothesis regarding the relation between subjective scores and behavioural metrics — the negative correlation between the score for 'AV understood me' and conflict in steering torques. More insight from the exploratory analysis are included in the discussions for each type of perception below.

Based on the qualitative data, *epistemic* concepts like intention and safety featured heavily in the factors affecting the perception of MHC concepts being evaluated. Additionally, moral and legal obligations, and untrustworthiness and novelty of the automation also played a crucial role in affecting the sense of responsibility of the participants. These high-level epistemic concepts were further dependent on low-level concepts relating to the trajectory of the vehicle and the nature of the interaction between the human and the automation.

Attributes of the *trajectory* like the timing of overtaking manoeuvres, distance to other vehicles, and direction of driving contributed to participant's perception of safety and understanding of the intentions of the AV. Matching intentions and driving styles of the AV positively affected the perception of MHC, while mismatches reduced the perception of MHC. Similarly, lack of safety negatively affected the perception scores.

The nature of the *interaction* between the driver and automation characterised by the torques, jerks and stiffness of the steering also influence the driver's perception of MHC. High torques, jerks or stiffness that increased the control effort, control overshoots, disruptions to fluent control and sudden changes in dynamics that warranted adaptations were cited as negatively affecting subjective perception of MHC. Subtle haptic guidance by the automation that aligned with the intentions of the driver had a positive influence on the perception of MHC, whereas, driver corrections due to mismatches in intent had a negative effect.

The following subsections provide a detailed discussion of the different factors affecting the perception of each MHC concept.

(1.) Sufficient control (MHC Property 3): None of our hypotheses regarding correlation of factors with the perception of 'sufficient control' were accepted in our confirmatory analysis. The exploratory analysis revealed that the score for 'sufficient control' was positively correlated with reaction time ($\beta = 0.24$, $p = 0.01$) and negatively correlated with trajectory deviation ($\beta = -0.19$, $p = 0.02$). And for HSC alone, the score for 'sufficient control' was positively correlated with conflict in steering torques ($\beta = 0.39$, $p = 0.01$) and overtaking time ($\beta = 0.35$, $p = 0.01$).

The most interesting finding from the exploratory analysis was the positive correlation of reaction time with the score for 'sufficient control', which contradicts our initial hypothesis. We had assumed that faster reaction times would have generated stronger perception of control. However, stronger perception of control might have made participants more complacent causing them to react late as found in previous studies (Payre et al., 2016; Dixit et al., 2016). This would mean that participants would have greater reaction times for the control mode for which they had greater sense of control. However, when comparing the control modes, HSC has a higher score for 'sufficient control' ($p < 0.001$) and a lower reaction time than TC ($p = 0.01$), and thereby contradicts the complacency explanation. Another plausible explanation might lie in the dynamics of the interaction - when participants react early, they only have to make smaller corrections than when they react late, which might lead to greater sense of control in cases where they react late. This presents an interesting contradiction where humans might be getting greater sense of agency for late reactions of greater magnitude than smaller reactions of small magnitude. Future research should explore this dynamic, especially in relation to sensory attenuation (where agents perceive their own actions to be of a smaller magnitude) and intentional binding (where agents perceive that the time interval between them doing some action and its outcome is smaller than the actual time elapsed) which have been related to sense of agency (Wen & Imamizu, 2022; Berberian et al., 2012b). To be more specific, sensory attenuation might cause them to feel that they are not doing much when agents are performing small actions when reacting fast (Bays et al., 2006). And intentional binding might be caused by extra cognitive load triggered when humans are being active (Wenke & Haggard, 2009), which might also explain why reaction times are higher when they are active.

AV understood me (MHC Property 2): The score for 'AV understood me' was negatively correlated with conflict in steering torques in accordance with our hypothesis ($\beta = -0.34$, $p = 0.001$). Furthermore, the exploratory analysis for TC alone revealed a positive correlation between time to collision (TTC) and the score for 'AV understood me' ($\beta = 0.25$, $p = 0.03$).

This perception had no statistical difference between HSC and TC. Quantitatively, the only metric that reached significance was *conflict*, which was negatively associated with "AV understood me." This is consistent with participants' qualitative reports that *force against their intention* or *invasive steering torque* undermined the feeling that the AV understood them. In other words, when drivers experienced strong resistance from the automation, they interpreted it as misalignment of intent. In the shared-control literature, such situations are often referred to as "fighting" (Abbink et al., 2012).

By contrast, *takeovers*, which we had hypothesised to be negatively related, showed near-zero association, and participants likewise did not mention takeover frequency as a reason for feeling misunderstood. Two non-hypothesised trends ran positive: higher *steering reversal rate* and greater *trajectory deviation* were weakly associated with stronger feelings that the AV understood the driver. Qualitative accounts help explain these patterns. Participants noted that when the AV *anticipated manoeuvre intentions* (e.g., lane alignment or well-timed overtakes), they felt understood. The plausible explanation of this is that trajectory deviations sometimes reflect the AV adapting to the driver's preferences rather than deviating from the ideal path. Similarly, participants described how *override capability* during negotiation supported their sense of being understood, aligning with the weak positive trend in steering reversal rate. In this view, deviation and reversal rate may reflect shared control responsiveness rather than poor control.

Thus, for "AV understood me," the most reliable behavioural metric was *low conflict*. This metric aligned with participants' accounts that forces against their intention indicated misunderstanding. Takeover counts provided little diagnostic value. Exploratory trends suggest that event-level *responsiveness* can foster perceived understanding. This responsiveness could be understood in the case of AV deviating from its nominal path to align with the driver, or making small co-directed adjustments. This highlights the value of developing event-based interaction measures rather than relying on global counts or RMS values.

I understood AV (MHC Property 2): The results did not indicate any correlations between the score for 'I understood AV' and the behavioural metrics. We might be tempted to say that the experiment might not have triggered a sufficient variation in the score for 'I understood AV', but, the significant difference in the scores for HSC and TC ($p < 0.001$) clearly negate this assumption. So, we need novel formulations of behavioural metrics for future experiments exploring the perception of 'I understood AV'.

Participants reported higher understanding of the AV under HSC than TC. Quantitatively, this perception was not significantly associated with any behavioural metric. All observed associations were non-significant, though some trended in expected directions. Conflict showed a negative trend, suggesting that less conflict may support better understanding, but the association was not statistically reliable. Reaction time and takeovers also trended in the expected directions: longer reaction times were weakly linked to greater understanding, whereas more frequent takeovers were weakly linked to lower understanding. By contrast, steering reversal rate, showed close-to-zero association, possibly because it was calculated as a root mean square (RMS) value across the whole drive rather than capturing event-specific corrections.

Qualitative findings provide context for the observed trends in conflict, reaction time, and takeovers. Participants reported that mismatched manoeuvre timing (too early or too late) and intrusive torque feedback undermined their ability to understand the AV. Similarly, erratic or unsafe manoeuvres that forced drivers to take over reduced perceived understanding, consistent with the weak negative trend in takeovers. At the same time, they noted that simply knowing how HSC and TC operated helped them to better interpret the AV's intentions, suggesting that system transparency and predictability play an important role in supporting this perception. Taken together, the results indicate that perceived understanding is less about overall input variability across a drive and more about specific moments of alignment or misalignment between the AV's actions and the driver's expectations. Future analyses should therefore prioritise event-based indicators (e.g., timing of manoeuvres, mismatched torque events) over global RMS metrics, which are too coarse to capture these dynamics.

Working together (MHC Property 2): Similar to 'I understood AV', the results for 'working together' did not indicate any significant correlations between the scores for 'working together' and any of the behavioural metrics even though there was a significant difference in the scores for HSC and TC ($p = 0.004$). Thus, novel formulations of behavioural metrics are needed for future research into the perception of the human 'working together' with the automation.

The final perception that differed significantly between HSC and TC was working together, with participants reporting higher values under HSC. This aligns with qualitative reports, where participants described HSC as fostering a greater sense of cooperation. Quantitatively, no behavioural metrics reached statistical significance. However, two out of four hypothesised metrics showed trends in the expected direction: more frequent takeovers were weakly associated

with lower ratings of working together, and greater conflict also trended negatively. Both patterns suggest that when drivers had to step in or felt "fought" by the automation, the sense of collaboration diminished.

Qualitative accounts reinforce this interpretation. Participants highlighted that subtle corrections, effortless overrides, and a sense of shared execution enhanced their feeling of working together. Conversely, poorly timed or misaligned manoeuvres, often experienced as either intrusive torque (conflict) or the need to take over, led participants to report that cooperation had broken down. By contrast, maximum steering torque and steering reversal rate showed no association with working together. One likely reason is that these were computed as RMS values across the whole drive, masking event-specific corrections that may matter most for this perception.

Trajectory deviation, although not part of our hypotheses, showed a weak positive trend with working together. Qualitative insights suggest a possible explanation: when the AV anticipated and executed manoeuvres (e.g., avoiding a bump), drivers often chose to follow its trajectory, reinforcing the sense of joint action. However, because such scenarios were relatively rare in our study, this trend should be interpreted with caution.

Overall, our results suggest that the perception of working together is not simply explained by how much or how often the driver applies physical input (e.g., maximum torque, reversal rate). Instead, it depends on the interactional qualities of those inputs, whether corrections feel subtle or forceful, whether overrides are smooth or resisted, and whether driver and AV actions are aligned in timing and intent. Relatedly, (Mars et al., 2014) emphasized the complementary perspective of conflict in cooperation, noting that to capture such moments of opposition (the inverse of working together), it is more informative to compute transient torque variations around specific events rather than relying solely on global measures averaged across the entire experiment.

Responsibility (MHC Property 4): None of our hypothesised correlations between the score for 'responsible' and behavioural metrics were accepted in the confirmatory analysis. The exploratory analysis for HSC alone, indicated a positive correlation between the score of 'responsible' and reaction time ($\beta = 0.18$, $p = 0.04$). And the exploratory analysis of TC alone revealed that the score for 'responsible' was negatively correlated with conflict ($\beta = -0.27$, $p = 0.03$) and positively correlated with maximum steering torque ($\beta = 0.24$, $p = 0.04$).

Statistical analysis showed no significant difference in perceived responsibility between HSC and TC, and no behavioural metrics were significantly associated with this perception. In our quantitative model, conflict showed a weak negative trend with responsibility, averaged across both modes. However, qualitative reports suggest that this association may differ by control mode: in HSC, high conflict reduced responsibility because participants felt forced to act, whereas in TC, low conflict reduced responsibility because cooperative execution diffused accountability.The LMM model may hide what's really going on, because conflict seems to have different effects in HSC and TC. In HSC, more conflict made drivers feel less responsible, while in TC, less conflict reduced responsibility. This suggests that responsibility is not just about the overall amount of conflict, but about how the conflict is shaped by the system's design.

Number of takeovers showed a weak positive trend, suggesting that frequent interventions may heighten the feeling of responsibility, though again not significantly. Qualitative accounts linked this to trust: participants who trusted the AV less felt more responsible and thus took over

more often, aligning with prior findings that higher trust reduces takeover frequency (Molnar et al., 2018). Context also mattered: in safe, low-demand driving, participants often felt less responsible, whereas in safety-critical scenarios, they felt heightened responsibility and were more likely to intervene.

Safety and trust: Besides the aforementioned perception scores related to MHC, we also recorded how participants perceived safety and trust. Safety (Peng et al., 2024; Chen et al., 2024; Papadimitriou et al., 2022) and trust (Payre et al., 2016; Dixit et al., 2016; Nordhoff et al., 2023; Molnar et al., 2018) are important qualities that have been extensively studied in the context of human-AV interactions. Even though safety and trust were not central to our exploration of MHC, we opted to include them in the hopes of informing research on safety and trust.

Our exploratory analysis revealed that the scores of 'safe' was negatively correlated with the jerk of vehicle trajectories ($\beta = -0.22$, $p = 0.01$) and positively correlated with overtaking time ($\beta = 0.22$, $p = 0.04$). Thus, our results support the findings of earlier works that posit that smoother vehicle trajectories would improve the sense of safety (Peng et al., 2024).

The exploratory analysis for the scores of 'trust' was positively correlated with trajectory deviation ($\beta = 0.24$, $p = 0.01$) and overtaking time ($\beta = 0.24$, $p = 0.05$). We would have expected that larger trust scores would result in the human driver intervening less, but the results indicate the opposite. A larger trajectory deviation is indicative of the human driver having more influence of the ego vehicle and this might indicate that drivers trust the system more when they can actively override it. Alternatively, trust might be influenced by other factors like safety - for example, trajectories with longer overtaking times might caused greater sensations of safety, which might in turn improve the perception of trust in the system. Future research could explore these complex relationships in greater detail.

It must be noted that overtaking time which is positively correlated with the scores for 'safe' and 'trust' was a metric that was tailored to the overtaking scenario in this experiment. In our experiment, since the potential collision happens when the ego vehicle leaves the overtaking lane prematurely and hits the motorcyclist beside it, longer overtaking times would be associated with greater perceptions of safety and trust. Thus, the results from our study might not generalise to another experiment where the potential collision happens in the overtaking lane. Thus, we advice future researchers to tailor the metrics and their hypotheses to each experiment design.

7.5.2 Comparing control modes

As shown in Figure 7.4(b), HSC supports MHC perception more than TC. Participants under HSC reported significantly higher scores for *sufficient control*, *I understood the AV*, and *working together*. Together, these map onto two properties of MHC, indicating that HSC better enables drivers to experience meaningful human control. Furthermore, *perceived safety* and *trust* were also rated higher for HSC compared to TC, even though they were not part of the core MHC perception evaluation.

In terms of behavioural outcomes, several differences between HSC and TC were consistent with our initial hypotheses. Specifically, participants under HSC showed faster reaction times, lower maximum steering torque, and more takeovers compared to TC, all aligning with the predicted directions in Table 7.2.These patterns suggest that HSC keeps drivers more engaged

and reduces the need for forceful corrective input. Other behavioural metrics, such as conflict, steering reversal rate, trajectory deviation, and jerk, showed no significant differences, leaving their role in supporting MHC less clear.

To further investigate how behavioural dynamics relate to perceptions, we analysed the significant associations between behavioural metrics and subjective ratings (Figure 7.5). For *sufficient control*, we had hypothesised a negative association with reaction time, expecting that longer delays to intervene would indicate diminished control in safety-critical situations. Instead, the observed association was positive: participants who intervened later reported stronger feelings of control. Further, trajectory deviation was also found to be negatively associated with sufficient control, even though this relationship was not hypothesised beforehand. For *AV understood me*, greater steering conflict was linked to lower ratings, consistent with the expectation that reduced "fighting" signals better alignment between driver and automation. By contrast, no behavioural metrics were significantly associated with *I understood the AV* or *working together*, despite their higher ratings under HSC.

Overall, the quantitative results indicates that HSC yields a more favourable MHC profile, with both perceptions and several behavioural metrics aligning with initial hypotheses. However, the factors underlying these perceptions, particularly the unexpected reaction time effect, remain to be clarified. We return to these mechanisms in the next section, drawing on qualitative evidence to explain how drivers experienced HSC and TC in practice.

7.5.3 Framework for Meaningful Human Control

Beyond comparing the two control modes HSc and TC, the findings from our study (Section sec:Framework-findings) can have broader implications for how humans perceive their interactions with automated systems. Based on our experiment, we propose guidelines for future studies of Meaningful Human Control (MHC) and propose a framework for designing experiments for evaluating MHC from the perspective of human operators interacting with an automated system.

Findings

Our evaluation framework contributes to the empirical assessment of Meaningful Human Control (MHC) in automated driving systems by translating the four MHC properties into observable indicators, linking these indicators to behavioural evidence, and offering a reproducible structure to model perceptions of MHC as a function of behavioural metrics and control mode.

First, we translated the four properties of MHC proposed by Cavalcante Siebert et al. (2023) into Likert-scale, self-reported quantitative questions. This extends prior work in several ways. For instance, Suryana et al. (2025b) evaluated MHC in the context of Tesla FSD Beta and Autopilot mainly through narrative descriptions of driving experiences, without questions explicitly designed to assess MHC. Our approach complements this by employing items that are intentionally constructed to measure the four properties. In parallel, Verhagen et al. (2024) advanced the field by proposing one of the first approaches to quantify MHC in human–robot teams through a 2D game-like firefighting simulation. Their focus, however, was primarily on the traceability condition, while leaving the tracking condition largely aside. In contrast, our

study is grounded in the four properties framework by Cavalcante Siebert et al. (2023) and uses a driving simulator that closely mirrors real driving tasks. Unlike the 2D firefighting game, which provides only a simplified representation, our setup offers an environment much closer to real-world conditions and thereby enables a more realistic evaluation of all four properties. Taken together, we see our contribution as an early attempt to operationalize MHC more comprehensively: not only by quantifying one dimension (e.g., traceability), but by systematically addressing all four properties.

Second, our evaluation results point to concrete design implications for driving automation systems that aim to comply with the four properties of MHC. *Property 1 (Moral Operational Design Domain (ODD))*: Drivers require system transparency and familiarisation to interpret automated behaviour. Training sessions, such as those we provided in our experiment, can help users understand system logic across routine and safety-critical scenarios without exposing them to real-world risks. Importantly, participants were explicitly informed that they remained fully responsible for the AV's actions and were required to stay aware at all times. Such reminders of accountability are essential for aligning drivers' understanding of their role with the moral ODD of the system. *Property 2 (sufficient control)*: Because higher trust can delay takeovers, systems should ensure that interventions remain easy and smooth, even after abrupt transitions. Continuous steering access, minimal override friction, and intuitive re-engagement mechanisms are critical design priorities. *Property 3 (shared representation)*: Our findings show that drivers evaluate automation at the event level rather than across entire trips. Thus, design should prioritise the quality of individual interactions—for example, ensuring subtle corrections, smooth overrides, and trajectory adaptations aligned with driver preferences. Aggregate smoothness metrics may miss these dynamics. *Property 4 (responsibility)*: Responsibility perceptions vary with system design and driving context. Designers should avoid extremes where drivers feel either forced by high conflict or disengaged through passivity. Interfaces and adaptive control strategies that sustain an appropriate sense of accountability, even in low-demand driving, represent a key area for future research. Together, these implications illustrate how our operationalisation of MHC can not only evaluate but also guide the design of shared-control systems.

Third, our LMM analyses illustrate a flexible analytical framework rather than a fixed equation. The strength of this approach lies in its structure: perceptions are modelled as a function of behavioural metrics and control mode, with participant-level random effects accounting for repeated measures. Researchers in other contexts can adapt the framework by selecting their own set of candidate metrics and control modes, tailoring models to the specifics of their study. In this way, our contribution is to provide a reproducible procedure for linking MHC perceptions with quantitative indicators, while allowing others to define the precise parameters of their own models.

Implications

Our findings have implications that extend beyond the specific comparison of HSC and TC, contributing to ongoing debates on complacency, sense of agency, and responsibility in human–automation interaction.

Complacency and sustained engagement. A recurring concern in Level 2–3 automation is that drivers may slip into passive monitoring roles, leading to complacency and delayed reac-

tions when manual intervention is required. Our results suggest that HSC may counteract this tendency: participants under HSC showed shorter reaction times and more frequent takeovers than under TC, indicating greater readiness to act. In contrast, TC, with its explicit but infrequent handovers, risks creating phases of overreliance during which drivers disengage until prompted. Our findings resonate with prior discussions on the two modes of control, where one of the identified pitfalls of TC is that drivers may fall into complacency (de Winter et al., 2023). At a societal level, these patterns underline the importance of designing shared control mechanisms that actively sustain driver engagement, rather than leaving humans in purely supervisory roles.

Sense of agency and quality of interaction. Prior research shows that higher levels of automation often reduce drivers' sense of agency (SoA), undermining engagement and willingness to accept responsibility. In our study, HSC elicited significantly higher ratings of *sufficient control*, *understanding the AV*, and *working together*. These perceptions map directly onto SoA, suggesting that continuous, transparent, and negotiated interaction helps drivers maintain a subjective feeling of authorship over vehicle behaviour. By contrast, TC can undermine SoA by producing disorientation during abrupt handovers or by diffusing the sense of joint action. Responsibility, however, showed no significant difference between HSC and TC, a point we elaborate in the next section, where we discuss how context and system design shape accountability.

Responsibility gaps and accountability. A persistent mismatch in automated driving is that while legal responsibility remains with the human driver, system design can diminish their felt responsibility. Our results reflect this tension. Quantitatively, responsibility ratings did not differ significantly between HSC and TC. Qualitative reports suggest that responsibility was shaped less by control mode itself than by situational demands and drivers' understanding of their operational role. For example, under HSC, some drivers felt *forced* into responsibility by counter-torque or abrupt interventions, whereas under TC, others felt disengaged or only partially responsible during cooperative execution. These findings highlight that responsibility is not fixed but shaped by how authority is balanced between human and automation: it can be reduced both when the system exerts too much pressure on the driver (dominance) and when it operates too independently, leaving the driver disengaged (independence). For designers and policymakers, bridging the gap between *assigned* and *experienced* responsibility is essential to prevent unfair attribution of blame.

Beyond driving. Although our study focused on automotive shared control, the broader implications extend to other domains where humans and automation jointly act, such as aviation, robotics, and healthcare. Across these domains, complacency, reduced agency, and blurred responsibility recur as central challenges. Meaningful Human Control offers a principled framework for addressing these challenges by guiding the design of systems that sustain engagement, preserve a sense of authorship, and maintain appropriate accountability. In this way, our work connects empirical findings on HSC and TC to broader societal debates on the responsible integration of automation.

Guidelines

Our work illustrates an example of converting high-level MHC concepts into subjective perceptions and objective metrics tailored to a particular scenario and testing the relationship between

these subjective perceptions and objective metrics. In Table 7.5, we summarise the steps in designing such an experiment. More experiments like this can be used to build a knowledge base of how different subjective perceptions are related to objective metrics. The proposed framework is thus useful for regulatory agencies and automation designers as described below.

Table 7.5: **Framework for designing MHC experiments:** When studying how humans interact with an automation system, the following steps can be used to design experiments for evaluating their subjective perceptions pertaining to meaningful human control and how they relate to behavioural metrics.

Step 1: Identify a scenario where there is potential conflict between the reasons of the human operator and the automation system which could be caused by mismatches in beliefs, desires or intentions.
Step 2: Identify the properties of MHC that are relevant to this scenario.
Step 3: Based on identified properties, identify subjective perceptions to be evaluated in the post-trial surveys and post-experiment questionnaires.
Step 4: Identify task-relevant objective metrics that pertain to the experiment scenario based on hypothesised relations to subjective perceptions.
Step 5: Conduct the experiment, collect data, and analyse the results.
Step 6: (Optional) Revise objective metrics and (questions for) subjective perceptions based on the participant responses to open-ended post-experiment questionnaire.

For regulating automated driving systems: The proposed methodology can be applied to design experiments for evaluating the performance and perception of human operators interacting with new designs of driving automation systems. In case of deficits, the lack of meaningful human control could be dealt with by requesting design updates or more training for operators.

For evaluating drivers and remote operators: For agencies engaged in the training and licensing of future drivers and remote operators, the proposed framework can be used to identify safety critical scenarios which can be used to test the performance of humans when interacting with automation systems.

For automation manufacturers: Experiments following this framework could be used to direct the design process and their results would be useful for justifying design choices in the face of law suits or regulatory scrutiny.

7.5.4 Limitations and future research

Level of guidance of shared control. In our experiment, the level of haptic guidance was fixed, even though shared control is not inherently binary. As (Mulder et al., 2012) emphasise, haptic shared control can vary continuously along a spectrum of guidance strength, characterised by the Level of Haptic Authority (LoHA). At low LoHA, the system provides only subtle guidance and the driver remains dominant; at high LoHA, the system exerts stronger torques and can

effectively drive the vehicle, approaching automation. Within this spectrum, TC corresponds to a high LoHA situation, whereas HSC represents an intermediate level where both driver and automation contribute. Our study therefore tested only two discrete points on this continuum, rather than exploring the broader range of LoHA values. Future work could adopt the experimental design by Mars et al. (2014), who analysed varying degrees of haptic control.

Participants. Most of our participants were students or individuals working in academia, many of whom had limited real exposure to automated vehicle technology. This sample may not fully reflect the perspectives of early adopters or more experienced users of automated technology. Future research should include a broader participant pool, incorporating drivers with more diverse ages, professional backgrounds, and driving experience.

AV outperforms Human. In this study, we explored a joint task where there is potential conflict between the human and the automation, where the human outperforms the automation (the AV fails to steer clear of some road users). Future research should also explore how MHC is affected in conflicting situations where the AV performs better than the human (like emergency braking or being aware of vehicles in the blindspot) - how the transfer of authority should be designed to account for such instances where AV oputperforms the human.

7.6 Conclusion

This research demonstrates the value of integrating telemetry data, structured questionnaires, and qualitative feedback to evaluate how drivers perceive meaningful human control in automated driving systems. This triangulated approach revealed that haptic shared control (HSC) supports meaningful human control better than traded control (TC), particularly in terms of perceived sufficient control, understanding the AV, and working together. However, no significance difference was observed in driver's perceived responsibility across the two control modes. Qualitative reflections indicated that driver's judgements of responsibility were more influenced by situational factors, such as trust in the system and the specific traffic context, rather than the control mode itself.

Our findings suggest that the perception of responsibility is complex and influenced by event-specific dynamics, such as the timing of manoeuvres and the degree of torque conflict, rather than aggregate metrics like trajectory deviation or jerk. This emphasises the importance of analysing system behaviour during concrete, real-time interactions rather than relying solely on aggregated data from the entire driving session. While HSC was more effective in supporting various MHC-related perceptions, perceived responsibility remained unaffected by control mode. This suggests that assigning responsibility solely to drivers may be insufficient unless complemented by making sure the driver's role and agency are preserved in the car and ensuring responsibility is fairly shared and not unfairly dumped on drivers.

In terms of design implications, our findings point to key strategies for improving MHC in partial driving automation systems: (1) provide training sessions for drivers, (2) ensure that interventions remain easy and smooth, (3) prioritise the quality of individual interactions, and (4) design interfaces and adaptive control strategies that sustain an appropriate sense of accountability. Future research should explore broader participant populations and scenarios that push the limits of driver-automation interactions to further validate and refine these insights.

Chapter 8

Conclusions and Perspectives

In this chapter, I present the overall conclusions and perspectives of this dissertation.

The objective of this dissertation to operationalise meaningful human control for automated-vehicle decision-making is addressed through seven research questions that are introduced in Section 1.6. Section 8.1 summarises the key findings associated with each research question. Section 8.2 provides the overall conclusion by integrating these findings and highlighting the main contributions of the thesis. Sections 8.3 and 8.4 discuss the implications for science and for engineering practice, and outline directions for future research.

8.1 Key Findings

This section provides direct answers to the research questions introduced in Chapter 1. For each question, a concise summary answer is first given, followed by an integrated elaboration based on the results across the relevant chapters.

RQ1: Which types of human reasons should automated vehicles prioritise when planning manoeuvres, and how should these priorities adapt to context-specific situations?

Automated vehicles should prioritise multiple interacting human reasons, with the safety of vulnerable road users as the highest priority, and adapt this prioritisation dynamically to contextual, regulatory, and situational conditions.

Through questions aimed at eliciting the reasons underlying automated-vehicle manoeuvre planning, a structured set of human reasons spanning four layers—normative, strategic, tactical, and operational—was identified. These reasons can be differentiated according to their temporal scale and the human agents to whom they are attributed. A single observable manoeuvre was consistently justified by multiple concurrent reasons, rather than by a single dominant reason. Moreover, the same reason was shown to operate differently across behavioural layers: for example, efficiency was interpreted as route optimisation at the strategic level but as smooth speed regulation at the tactical level. In addition, contextual factors such as local regulations, technological capabilities, levels of automation, traffic situations, and individual differences systematically influenced how reasons emerged and were interpreted. The overtaking-cyclist case study further demonstrates how the safety of vulnerable road users assumes the highest priority, while allowing context-sensitive flexibility in rule compliance, illustrating how prioritisation can directly shape decision logic in everyday driving when facing ethically challenging situations.

RQ2: How can human reasons be represented in a form suitable for supporting automated-vehicle decision making?

Human reasons can be represented through structured relationships between underlying reasons and preferred automated-vehicle behaviour, informed by human-factors insights.

The findings show that insights from human-factors research provide a connecting layer between abstract human reasons and expected automated-vehicle behaviour in concrete traffic contexts. These relationships clarify how specific behaviours in particular situations are interpreted as expressions of underlying reasons from the human perspective. When interpreted through an engineering lens, such relationships can be translated into formalised representations suitable for use in computational frameworks. Although these representations are not intended as universal descriptive models, they demonstrate how human-factors knowledge can support the translation of human reasons into supervision and evaluation mechanisms within automated-vehicle systems.

RQ3: How can vehicle control frameworks integrate human reasons to enable dynamic behavioural adaptation when misalignment occurs?

Human reasons can be integrated into vehicle control frameworks through supervisory evaluation mechanisms layered on top of existing controllers.

By converting alignment with human reasons into a continuous evaluative indicator, automated vehicles can assess in real time whether current states and candidate trajectories remain consistent with the reasons of relevant human agents. This evaluative component can be embedded as an auxiliary supervisory layer rather than requiring redesign of low-level controllers. When alignment falls below a specified threshold, adaptive re-planning is triggered, allowing the system to restore consistency between behaviour and human reasons while maintaining safety, legality, and efficiency. This establishes a closed supervisory loop between normative evaluation and motion planning.

RQ4: How can AV trajectories be systematically evaluated and selected based on relevant human reasons to satisfy the tracking condition of meaningful human control?

AV trajectories can be systematically evaluated using explicit reason-based weighting and balance mechanisms that ensure transparent prioritisation across multiple human agents.

The trajectory-evaluation mechanism demonstrates how conflicting reasons across agents can be reconciled through explicit weighting. In safety-critical interactions, the safety of vulnerable road users consistently received the highest weight, followed by rule compliance and efficiency. To prevent systematic exclusion of any agent's reasons, a balance function was introduced to penalise zero-weight solutions. Mathematical analysis showed that maximum balance occurs when all relevant agents exert non-zero influence. When integrated into trajectory evaluation, this function ensures proportional representation of all relevant human reasons and supports transparent, fairness-aware trajectory selection.

RQ5: How do drivers' perceptions of safety and trust relate to the extent to which automated vehicles track their reasons?

Drivers' perceived safety and trust generally increase when automated-vehicle behaviour aligns with their safety-related tactical and operational reasons, but this relationship is conditional: misalignment in clearly dangerous situations reliably reduces trust and perceived safety, whereas benign deviations do not necessarily do so.

When automated behaviour followed expected safety strategies—such as smooth braking, sufficient spacing, and prudent lane changes—drivers consistently reported higher trust and perceived control. However, not all forms of misalignment led to negative perceptions: failures to track expected lane changing or braking behaviour were not necessarily associated with reduced safety or trust when the situation was not perceived as dangerous. In contrast, misalignment in clearly safety-critical situations consistently resulted in low trust and perceived lack of safety. These findings indicate that perceived safety can serve as a partial behavioural proxy for whether the system tracks the human reason of safety, but that important discrepancies between perceived safety and actual tracking remain.

RQ6: How can drivers' subjective experiences be used to evaluate whether automated vehicles operate under meaningful human control?

Drivers' subjective experiences—expressed through perceived safety, trust, responsibility, and readiness to intervene—provide qualitative indicators for evaluating both tracking and tracing under meaningful human control in partially automated driving systems.

Subjective interview data enabled the evaluation of both tracking and tracing in real-world partially automated driving. Tracking was assessed through drivers' perceptions of whether safety-related subsystems consistently performed their intended functions, while tracing was

evaluated through drivers' awareness of supervisory responsibility, readiness to intervene, and understanding of their moral accountability. The results show that while partial adherence to meaningful human control was observed, significant inconsistencies remain. Tracking weaknesses were linked to technological limitations such as false positives, false negatives, and sensor vulnerabilities, as well as to misaligned user expectations. Tracing was found to fluctuate with perceived risk: during routine low-risk operation, drivers' perceived responsibility and readiness to intervene declined, whereas in higher-risk contexts, vigilance and perceived responsibility increased. These findings demonstrate that both tracking and tracing are dynamically shaped by system performance, user expectations, and situational risk, revealing a structural vulnerability in current partial automation paradigms that depend on sustained human oversight.

RQ7: How can vehicle telemetry data be used to evaluate whether automated vehicles operate under meaningful human control?

Vehicle telemetry can be used to evaluate meaningful human control when analysed at the level of concrete interactions and triangulated with subjective perceptions, allowing tracking to be inferred from behavioural dynamics, while tracing requires additional contextual and experiential interpretation.

Objective behavioural metrics, including steering torque, torque conflicts, reaction times, and intervention frequency, were shown to be closely linked to drivers' perceived control and understanding of the automated vehicle. Crucially, perceptions of meaningful human control depended more strongly on event-level interaction dynamics, such as the timing of manoeuvres and moments of haptic conflict, than on aggregate trip-level metrics. By integrating telemetry data with structured questionnaires and qualitative feedback, tracking could be evaluated through observable system–driver interaction patterns, while tracing required interpretation of drivers' perceived agency, accountability, and situational trust. The results further show that perceived responsibility did not vary systematically with control mode and could not be reliably inferred from telemetry alone, but instead depended on contextual factors such as trust and traffic conditions. Together, these findings establish that telemetry provides a necessary but not sufficient basis for evaluating meaningful human control and must be combined with subjective measures to assess tracking and tracing in real and simulated environments.

8.2 Overall Conclusions

This section synthesises the scientific conclusions of the dissertation across the ethics, engineering, and evaluation aspects of meaningful human control. Rather than reiterating individual chapter results, the conclusions articulate what is now known that was not known before, how these findings relate to prior work on meaningful human control and automated-vehicle ethics, and to what extent the three research gaps identified in Chapter 1 have been addressed.

Ethics Aspect: Representation and Prioritisation of Human Reasons

This thesis establishes that the human reasons guiding automated-vehicle manoeuvre planning are inherently multi-layered, context-dependent, and often simultaneous, and therefore cannot

be adequately captured by fixed or purely abstract prioritisation schemes. Prior philosophical accounts of meaningful human control proposed that automated vehicles should track both proximal and distal human reasons and resolve conflicts through general prioritisation rules (Mecacci & Santoni de Sio, 2020). However, these accounts remained largely normative and were illustrated primarily through hypothetical scenarios. The present findings complement and extend this work by providing the first empirically grounded specification of how human reasons are represented and prioritised in everyday automated-driving contexts.

With respect to representation, the results show that human reasons can be structured as relationships between underlying normative motivations and preferred automated-vehicle behaviours across normative, strategic, tactical, and operational layers. A single observable manoeuvre was consistently justified by multiple concurrent reasons rather than by a single dominant objective, and the same reason (e.g., efficiency or safety) took on different behavioural meanings across layers. These findings demonstrate that human reasons cannot be represented as isolated values or fixed cost terms but must instead be represented as context-sensitive relations between reasons and expected vehicle behaviour. This provides an empirically grounded basis for translating philosophical notions of human reasons into structured design knowledge suitable for supervisory and evaluative system components.

In terms of prioritisation, the findings show that reason hierarchies are not fixed but adapt systematically to regulatory environments, traffic situations, technological capabilities, and individual expectations. The overtaking-cyclist case study further refines existing MHC theory by showing that prioritisation is not governed solely by abstract notions of proximity but is shaped by context-specific vulnerability and risk exposure. Across expert reasoning, the safety of vulnerable road users consistently emerged as the highest-priority consideration, while rule compliance and other secondary values remained important but conditional. This provides an empirically grounded prioritisation principle for ethically routine driving situations, extending earlier hypothetical formulations of reason responsiveness.

Taken together, these findings close the previously identified ethics gap in MHC research by moving from abstract reason taxonomies toward empirically grounded representations and context-sensitive prioritisation principles for automated-vehicle decision-making. Human reasons are shown not merely to exist at different conceptual levels but to form a structured, situationally adaptive basis for determining what automated vehicles ought to do in specific traffic interactions.

Engineering Aspect: Embedding Human Reasons into Control Frameworks

Prior engineering work relevant to meaningful human control has not yet provided a systematic and operational method for embedding human reasons into standard automated-vehicle (AV) control frameworks in a way that enables transparent trajectory evaluation and dynamic behavioural adaptation. Although Calvert & Mecacci (2020) articulated the conceptual need for reason-responsiveness, their proposal did not specify how such responsiveness could be instantiated within real-time planning and control architectures. Related work on ethical and value-based control has demonstrated how abstract values may influence optimisation-based controllers through cost functions or rule hierarchies (e.g., Thornton et al., 2016, 2018; Geisslinger et al., 2023), but without making the underlying reason-selection and weighting procedures explicit at the trajectory-evaluation level. As a result, prior approaches have not resolved the

implementability and transparency limitations identified in the engineering gap.

This thesis advances the field by demonstrating that human reasons can be integrated into existing AV control frameworks through a supervisory evaluation layer, rather than by constructing new controllers. This represents a shift from reasoning embedded implicitly within objective functions to a modular supervisory mechanism that evaluates trajectories according to their alignment with human reasons. The result is a system in which human-reason responsiveness becomes a computational property of the overall architecture, not a redesign of the controller itself.

Three technical developments underpin this conclusion:

1. Reason alignment can be converted into a continuous evaluative metric. This provides an operational method for assessing whether a planned trajectory remains consistent with the reasons of relevant human agents. This was previously absent from both ethical control literature and the MHC framework, which defined reason-responsiveness normatively but did not specify an implementable metric.

2. The supervisory layer can trigger adaptive re-planning when misalignment occurs. This establishes a closed loop between normative evaluation and motion planning. Earlier ethical frameworks typically evaluated behaviour post hoc or at design time; they did not regulate behaviour dynamically. This implementation shows that AV decision-making can remain responsive to human reasons while retaining compatibility with existing MPC-based planners.

3. Conflicting stakeholder reasons can be balanced explicitly using a structured weighting and fairness mechanism. Prior formulations of MHC acknowledged the existence of multiple agents' reasons but did not operationalise how conflicts should be handled. Introducing a non-zero fairness constraint ensures that each stakeholder's reasons contribute to trajectory evaluation, thereby preventing systematic exclusion and maintaining procedural fairness.

Together, these developments close the engineering gap by demonstrating that reason responsiveness is technically implementable as a supervisory integration problem rather than a controller design problem. This reframes the engineering challenge of MHC: the central task is not to build controllers that directly "reason", but to design supervisory structures that ensure existing controllers behave in ways that remain aligned with human reasons.

Evaluation Aspect: Empirical Assessment of Tracking and Tracing

Prior work on meaningful human control has primarily addressed tracking and tracing at a conceptual level or through post hoc responsibility analysis following accidents (Calvert et al., 2020b, 2021). As a result, it remained largely unknown how the tracking and tracing conditions of MHC manifest during real-time interaction between drivers and automated vehicles, how they are perceived by users, and to what extent they can be evaluated empirically during ongoing operation.

This thesis establishes that meaningful human control can, in fact, be empirically assessed in partially automated driving through a combined analysis of drivers' subjective experiences and

objective vehicle telemetry. Rather than being purely theoretical properties of system design, tracking and tracing are shown to be dynamically expressed in driver–vehicle interaction.

Regarding tracking, the findings show that drivers' perceived safety and trust in real-world, partially automated driving systems are conditionally sensitive to whether automated-vehicle behaviour aligns with their safety-related reasons. When automated behaviour followed expected safety strategies, drivers consistently reported higher trust and perceived safety. Conversely, misalignment in clearly safety-critical situations reliably degraded trust and perceived safety. However, benign deviations in non-critical contexts did not necessarily lead to negative perceptions. These results demonstrate that perceived safety functions as a partial behavioural proxy for safety-related reason tracking, but important discrepancies between perceived safety and actual behavioural alignment can persist. This complements previous findings by Calvert et al. (2020b), which indicate that the vehicle's ability to track safety reasons was not consistently fulfilled across the operation of partially automated driving systems. However, it also shows that these conditions are dynamic, with some situations fulfilling them and others not, as situational risk moderates their fulfilment. Further, it shows that perceived safety and trust could be used as a proxy to evaluate safety-related reasons, but this is only applicable when there is alignment or misalignment in safety-critical situations; in non-critical situations, misalignment may still be perceived as tracking.

With respect to tracing, the findings demonstrate that drivers' awareness of responsibility, readiness to intervene, and perceived moral accountability vary dynamically with context and perceived risk. During routine, low-risk automated operation, supervisors' vigilance and perceived responsibility systematically declined, while higher-risk contexts reactivated supervisory awareness and intervention readiness. Tracing therefore emerges not as a static property guaranteed by system design alone, but as a dynamic human–system relation that fluctuates with system performance, user expectations, and situational risk. This provides the first empirical demonstration of a structural vulnerability in partial automation with respect to sustained tracing, which had previously been assumed rather than tested in MHC theory.

In terms of objective assessment, driving behavioural telemetry can infer tracking at the level of concrete interaction events but not tracing in isolation. Measures such as steering torque, torque conflicts, reaction times, and intervention patterns were closely linked to drivers' perceived control and trust, enabling misalignment to be detected through behavioural dynamics. However, perceived responsibility and moral accountability could not be reliably inferred from telemetry alone and required contextual interpretation. These results confirm that telemetry is necessary for evaluating meaningful human control, but it is not sufficient on its own. This is consistent with the views expressed by Verhagen et al. (2024), who argue for the combination of subjective and objective metrics to provide a comprehensive assessment of MHC.

Taken together, these findings close the previously identified evaluation gap in MHC research by demonstrating that tracking and tracing can be empirically assessed during ongoing automated driving, but only through a multi-layer evaluation framework that integrates subjective perception and telemetry data. In doing so, the thesis extends MHC from a predominantly normative and post hoc evaluative concept to an empirically tractable property of real-time human–automation interaction.

8.3 Scope and limitations of the current work

The scope of this dissertation is the operationalisation of meaningful human control for automated-vehicle decision-making, with a primary focus on the tracking condition. The work demonstrates how human reasons can be formalised, embedded within supervisory and evaluation mechanisms, and examined empirically through simulated and real-world studies. The tracing condition is considered at the level of evaluation rather than as an integrated component within control frameworks, and therefore remains only partially addressed in the present implementation.

While the framework captures core aspects of meaningful human control, several limitations constrain its current applicability. The implementation is limited by the availability of data that represent human reasons with sufficient granularity, and the evaluation relies on simulated and small-scale user studies. Broader validation is needed across more diverse user groups, driving cultures, and technological deployments to strengthen empirical generalisability. Furthermore, the catalogue of human reasons was derived mainly from experts within Western contexts and may not fully represent culturally diverse expectations. Relatedly, the human agents consulted in Chapter 2 reflect expert perspectives rather than the preferences of the broader public; incorporating public preferences, through methods such as discrete choice modelling applied to the general public, falls outside the scope of this thesis but represents a complementary direction directly relevant to regulators seeking to align AV behaviour with broader public values Gros et al. (2025a).

From a methodological perspective, the present trajectory-evaluation and supervisory framework addresses a relatively simple overtaking scenario with limited environmental uncertainty. More complex traffic environments, such as intersections, merging, or interactions with multiple heterogeneous road users, remain beyond the tested scope. The current mathematical representation models human reasons using a simple exponential decay function; future empirical work will be required to infer more realistic functional forms and to validate parameterisation based on experiment.

Finally, empirical findings highlight substantial variation in how individuals interpret safety-related behaviour, indicating that a single fixed model of human-reason representation cannot fully capture the diversity of preferences expressed by different stakeholders. This variability suggests the need for adaptive or personalised reasoning models that respond to individual expectations in addition to shared values.

The following sections discuss the scientific and practical implications of these findings and outline directions for future research toward scalable and real-time applications.

8.4 Implication for practice

The framework developed in this thesis demonstrates how human reasons can be operationally embedded into automated-vehicle decision evaluation through a supervisory mechanism that monitors behavioural alignment and triggers adaptive replanning when misalignment occurs. By extending evaluation beyond purely safety- or performance-based criteria, the framework addresses the challenge of unreasonable behaviour in ethically routine driving situations. In

practical terms, it shows how reason-based supervision can complement existing control architectures by adding a layer that explicitly evaluates decisions in terms of their alignment with prioritised human reasons.

The framework further illustrates how reason-based evaluation could conceptually complement existing assessment benchmarks, such as the United Nations Economic Commission for Europe's notion of the "competent and careful driver," by extending attention from safety-critical events to ethically ambiguous everyday situations. It also shows how the Safety of the Intended Functionality (SOTIF) perspective could be broadened to include evaluations based on human-reason descriptions rather than solely on telemetry-derived performance indicators. The alignment metric and associated evaluation tools produced in this work therefore provide a conceptual template for assessing tracking under meaningful human control, while not yet constituting a validated regulatory instrument.

For **system designers and developers**, this research provides concrete guidance on how human-reason responsiveness can be incorporated into existing automated-vehicle architectures without requiring fundamental redesign of low-level controllers. The proposed human-reasons-based supervision and trajectory-scoring components can be implemented as a modular layer connected to the global planner, enabling continuous monitoring of system behaviour and adaptive replanning when misalignment with human reasons is detected.

To practically implement such a module, developers require access to vehicle state data, candidate trajectories, stakeholder-specific weighting schemes, and models that relate perceived safety and trust to observable behavioural indicators. The findings suggest that design attention should focus in particular on safety-critical interactions involving vulnerable road users, where misalignment has the strongest impact on trust and perceived control. At the same time, the work highlights that reason-based models currently remain scenario-specific and require broader experimental validation. Further testing in a wider range of manoeuvres and traffic contexts is therefore necessary, as is careful analysis of controller stability under repeated supervisory interventions.

For **regulators and policymakers**, the framework illustrates how evaluations of automated driving could be conceptually extended beyond compliance with functional safety and performance requirements toward assessments of ethical and societal alignment. In principle, the alignment metric could be used as one input into supervised testing procedures to examine whether automated vehicles tend to prioritise human reasons in ways that are consistent with regulatory and societal expectations.

However, the results of this thesis do not support direct regulatory deployment of the framework in its current form. Threshold values for acceptable alignment, the selection and weighting of stakeholder reasons, and the interpretation of alignment scores necessarily involve normative judgement and expert oversight. Any future regulatory use would therefore require standardisation, large-scale validation, and institutional agreement on how reason-based evaluation should complement existing type-approval procedures rather than replace them.

Beyond technical and regulatory design, the findings also have direct implications for **responsibility and governance** in partially automated driving. The empirical results show that drivers' perceived responsibility and readiness to intervene decline during routine automated operation and increase primarily in situations perceived as risky. This finding indicates that reliance on continuous human vigilance as the primary safeguard for meaningful human control

is structurally fragile.

Responsibility for automated-vehicle behaviour should therefore be distributed across the socio-technical system rather than placed solely on the human driver. Developers remain responsible for the behavioural logic and supervisory assumptions embedded in the automation, including how transitions between manual and automated modes are structured. Regulators, in turn, bear responsibility for specifying acceptable operational domains and levels of automation that remain compatible with human cognitive and attentional capabilities. A shared governance model that explicitly recognises the limits of human supervision while reinforcing accountability across design, deployment, and oversight is therefore essential for sustaining meaningful human control in practice.

8.5 Implication for science and recommendation

This thesis advances the scientific understanding of meaningful human control (MHC) by moving beyond its predominantly conceptual treatment in the literature towards a systematic technical and empirical operationalisation for automated-vehicle decision-making. Whereas prior work on MHC has primarily focused on philosophical clarification and normative design principles (e.g., (Santoni de Sio & Van den Hoven, 2018; Mecacci & Santoni de Sio, 2020), this research demonstrates how human reasons can be represented, embedded in supervisory control, and empirically assessed using both behavioural and telemetry-based data. In doing so, it contributes to bridging a persistent divide between ethical theory, human-factors research, and automated-vehicle control engineering.

From a human–automation interaction perspective, this thesis provides empirical evidence that perceived safety, trust, responsibility, and readiness to intervene are systematically related to tracking and tracing under MHC. This extends prior trust and shared-control research by embedding these constructs within a formal ethical-control framework. From a control and decision-making perspective, the work demonstrates that ethical reasoning need not remain external to vehicle control architectures but can be implemented as a computational supervisory layer that evaluates and adapts behaviour in real time. Collectively, these results establish MHC not only as a normative requirement, but as a measurable and engineerable system property.

Building on these scientific implications, the following recommendations identify concrete directions for future research that arise directly from the limitations and findings of this thesis.

- *Empirical Validation of Human-Reason Models*

  The current mathematical representations of human reasons are theoretically grounded but require systematic empirical calibration. The proposed decay functions that describe how alignment with human reasons diminishes under behavioural deviation were chosen for conceptual clarity rather than empirical optimality. Controlled experiments combining behavioural, physiological, and subjective measures could be used to estimate these functions and validate threshold parameters. Such validation would strengthen the role of reason-alignment metrics as scientifically grounded rather than purely normative constructs.

- *Cross-Cultural and Contextual Studies*

The human-reason taxonomy and prioritisation structures developed in this thesis are primarily derived from expert input within Western regulatory and cultural contexts. However, prior research in traffic psychology and AV acceptance suggests that values, risk tolerance, and interaction norms vary significantly across societies. Comparative cross-cultural studies could therefore refine which elements of human-reason alignment are universal and which are context-specific. This would be essential for transferring MHC-based supervision frameworks across legal systems and traffic cultures.

- *Expansion to Complex Traffic Scenarios*

  The present empirical and simulation-based validations focused primarily on cyclist overtaking and routine automated driving scenarios. While these cases capture ethically salient everyday conflicts, future work should extend the framework to more complex interaction settings, such as unsignalised intersections, multi-agent merging, pedestrian-dense urban environments, and emergency manoeuvres. Evaluating the supervision and trajectory-scoring mechanisms under such conditions would test the scalability and robustness of the operationalisation of MHC under higher interaction complexity.

- *Evaluation Using Open Datasets*

  This work shows that behavioural data alone are insufficient to directly reveal the human reasons underlying automated-vehicle decisions, but also illustrates how behavioural patterns can be interpreted through a human-reason lens. Large-scale open datasets such as Waymo or nuScenes provide a unique opportunity to examine which implicit value structures are embedded in production-level AV behaviour. Applying the proposed supervision framework retrospectively to such datasets could enable large-scale empirical audits of tracking under MHC and reveal systematic misalignments between claimed design intentions (e.g., safety, comfort) and observable behaviour.

- *Integration into Global Planner and Control Algorithms*

  In its present form, MHC is implemented as an external supervisory layer that evaluates and adapts outputs of an existing global planner. Future research could explore deeper architectural integration, for instance by encoding human reasons directly as objectives or constraints within optimal-control, sampling-based planning, or game-theoretic formulations. Comparing external supervision with internally reason-aware planners under identical scenarios would clarify the trade-offs between modularity, transparency, stability, and real-time performance.

- *Implementation of Human Reasons in Foundation Models*

  Recent advances in foundation models and large-scale learning-based reasoning systems raise the question of whether such models could support automated-vehicle decision-making at higher cognitive levels. Although their internal reasoning processes do not necessarily reflect human reasons in a normative sense, future work could investigate whether they can generate interpretable candidate explanations for AV actions that can be evaluated by the human-reasons supervision layer. This could provide a hybrid architecture in which data-driven inference is constrained and audited by explicit MHC principles.

- *Integration with Control and Learning Systems*

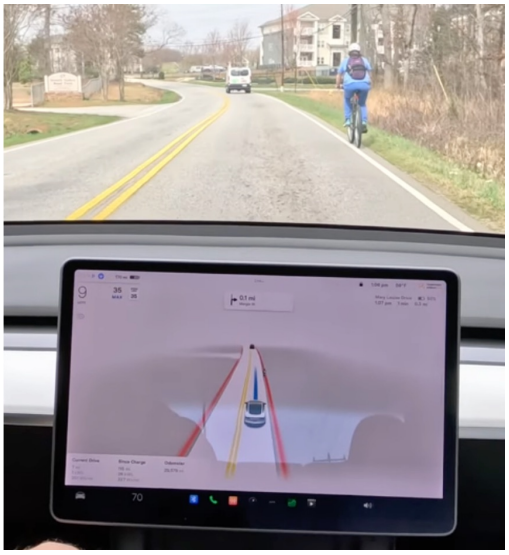
  A further promising direction is the integration of human-reason supervision into learning-based or probabilistic controllers. Many contemporary AV systems rely on reinforcement learning, imitation learning, or uncertainty-aware Bayesian approaches. Embedding MHC-based supervision within such controllers could support ethical adaptivity under uncertainty while preserving transparency and accountability. This would directly address current concerns in the literature that learning-based AV systems remain difficult to govern ethically due to their opacity.

Appendix

A Appendix for Chapter 2: Reasons and Principles for Automated Vehicle Manoeuvre Planning

A.1 Questionnaire

Table 1: Interview questions

Question number	Question
Q1	Where do you work, what is your position and role, what are your activities with regard to automated vehicles (AVs)?
Q2	What should automated vehicles (AVs) consider when planning a maneuver? Please give one example in much detail as possibl
Q3	Which moral aspects do you believe AVs should consider when planning a maneuver?
Q4	How might these aspects affect the manoeuvre plan?
Please watch the video below and read its description  **Video description** A passenger uses an automated vehicle (AV) for a morning commute to the office. The passenger has an important meeting and must arrive on time. If the vehicle maintains the current speed, the passenger can reach the office on time in 20 minutes. The AV is on a road with solid double yellow lines, which prohibit vehicles from crossing in both directions due to safety reasons. During the trip, the AV approaches a cyclist traveling at half of the speed of the AV. There is no safe passing zone visible from the vehicle; however, the opposite lane is currently empty.	
Q5	If the video continues, what do you believe all traffic participants will do?
Q6	What are the reasons for the [traffic participants mentioned by the experts] performing the [actions the experts mentioned]?

Continued on next page

Table 1 – Continued from previous page

<table>
<tr><th>Question number</th><th>Question</th></tr>
<tr><td>Q7</td><td>Besides the [traffic participants that are mentioned by the experts]'s, can you think any other factors that might influence the traffic participant decisions?</td></tr>
<tr><td>Q8</td><td>What do you think the reasons are for the [other factors that are mentioned by the experts]?</td></tr>
<tr><td>Q9</td><td>Can you think of any situations where the intentions of the [traffic participants / other factors the experts mentioned] might conflict? Please share any examples you can think of, and let me know when these conflicts may typically occur.</td></tr>
<tr><td colspan="2">Recall the scene from the previous video.
There are three different people, each with their own intentions:
• The automated vehicle (AV) passenger wants to pass the cyclist to get to the office on time.
• The cyclist wants a safe distance from the AV for safety concerns.
• The road policymaker wants both AV and cyclist to use their designated lanes, marked by solid yellow lines, for everyone's safety.
Keep this in mind as you answer the rest of the questions.</td></tr>
<tr><td>Q10</td><td>From your perspective, whose intentions should be given the most importance? Please answer this question by ranking the individuals below, with '1' indicating the highest rank.
<table>
<tr><th></th><th>1</th><th>2</th><th>3</th></tr>
<tr><td>AV passenger</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Cyclist</td><td>◯</td><td>◯</td><td>◯</td></tr>
<tr><td>Road policymakers</td><td>◯</td><td>◯</td><td>◯</td></tr>
</table></td></tr>
<tr><td>Q11</td><td>Could you please explain the reasons behind the rank you provided in your previous answer?</td></tr>
</table>

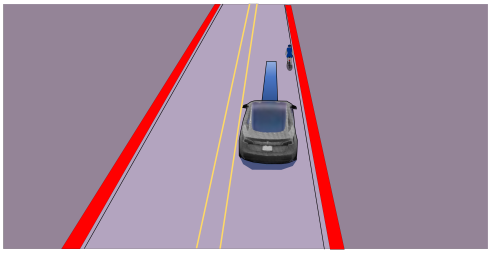
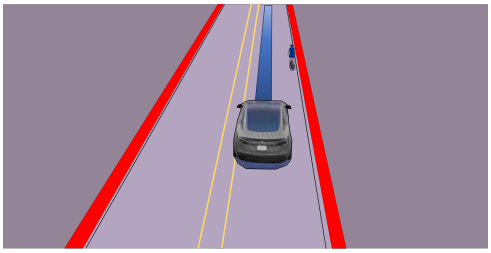
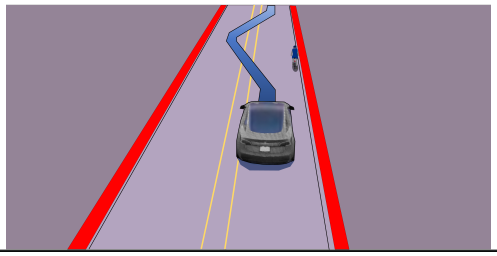
Continued on next page

Table 1 – Continued from previous page

Question number	Question
	Watch the three video scenarios below! These scenarios show three possible actions the AV might take if the previous video continues. The blue line ahead of the AV indicates the path it will follow. Imagine the AV's speed is the same in scenarios 2 and 3. **Scenario 1: AV stays behind the cyclist** In this scenario, the AV only considers the cyclist's need for a safe distance and the road rules that require the AV to stay in its lane. But it doesn't consider the AV passenger's desire to get to the office on time. **Scenario 2: AV overtakes the cyclist on its own lane** In this scenario, the AV is solely concerned with the AV passenger's goal of getting to the office on time and the road rules that insist on it staying in its lane. But it doesn't consider the cyclist's wish to ride with a sense of safety. **Scenario 3: AV overtakes the cyclist by using the opposite lane** In this scenario, the AV is focused on the AV passenger's concern about getting to the office on time and the cyclist's concern about a safe distance. But it doesn't consider the road rules that require it to stay in its own lane.
Q12	Which of the above scenarios do you prefer? Please answer this question by ranking the scenarios, with '1' indicating the highest preference.

Continued on next page

Table 1 – Continued from previous page

Question number	Question
	 Scenario 1: AV stays behind the cyclist  Scenario 2: AV overtakes the cyclist on its own lane  Scenario 3: AV overtakes the cyclist by using the opposite lane
Q13	Could you please explain the reasons behind the rank you provided in your previous answer?

Continued on next page

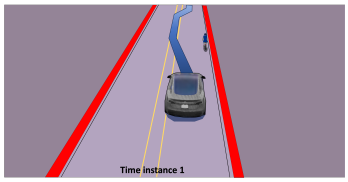
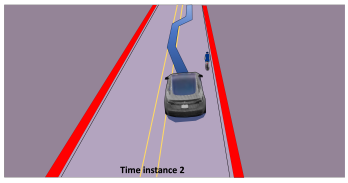
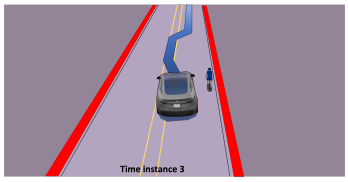
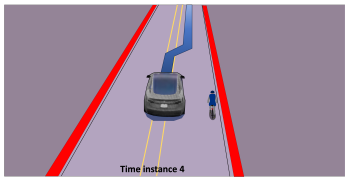
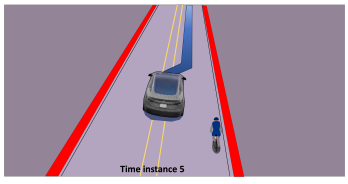
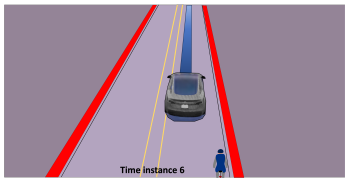
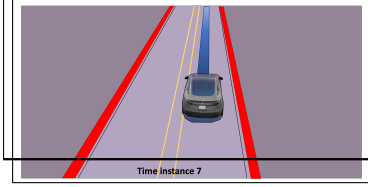
Table 1 – Continued from previous page

Question number	Question

Take a look at the video below!

Scenario 4: AV overtakes the cyclist by crossing some part of the opposite lane

In this scenario, the AV only considers the cyclist's need for a safe distance and the road rules that require the AV to stay in its lane. But it doesn't consider the AV passenger's desire to get to the office on time. You will now assess how much you believe the AV considers the intentions of three different stakeholders across 7 moments in this scenario. The same response table below will be used to answer Questions 14, 16, and 18.

	0	10	20	30	40	50	60	70	80	90	100
Time instance 1 	○	○	○	○	○	○	○	○	○	○	○
Time instance 2 	○	○	○	○	○	○	○	○	○	○	○
Time instance 3 	○	○	○	○	○	○	○	○	○	○	○
Time instance 4 	○	○	○	○	○	○	○	○	○	○	○
Time instance 5 	○	○	○	○	○	○	○	○	○	○	○
Time instance 6 	○	○	○	○	○	○	○	○	○	○	○
Time instance 7 	○	○	○	○	○	○	○	○	○	○	○

Continued on next page

Table 1 – Continued from previous page

Question number	Question
Q14	Imagine you are the AV passenger in scenario 4. Please assess, for each time instance, how much you believe the AV considers your intention to arrive at the office on time. (0 = Not consider at all; 100 = Fully consider)
Q15	Please clarify why the scores you provided for each time instance are either constant or change at each time instance
Q16	Imagine you are the cyclist that is passed by the AV in scenario 4. Please assess, for each time instance, how much you believe the AV considers your intention to bike with a sense of safety. (0 = Not consider at all; 100 = Fully consider)
Q17	Please clarify why the scores you provided for each time instance are either constant or change at each time instance
Q18	Imagine you are a road policymaker, and you see the AV briefly occupying small part of the opposite lane when overtaking the cyclist in scenario 4. Please assess, for each time instance, how much you believe the AV considers your intention putting a solid double yellow lines on the road for safety reasons. (0 = Not consider at all; 100 = Fully consider)
Q19	Please clarify why the scores you provided for each time instance are either constant or change at each time instance
Q20	How can an AV understands the intention of people on the road and other stakeholders in a real traffic situation?
Q21	From your point of view, what approach should AVs use to manage possible conflicts between the intentions of people on the road and other stakeholders in real traffic situations?

B Appendix for Chapter 7: Subjective Assessment of Meaningful Human Control

Table 2: Interview questions

Question number	Question
Q1	Do you have the Full Self-Driving Beta (FSD Beta) feature? (1 = Yes, 2 = No)
Q2	Before the first time of using Autopilot and FSD Beta, did you watch / read / listen to information on how to use it? (1 = Yes, 2 = No)

Continued on next page

Table 2 – Continued from previous page

Question number	Question
Q3	Please mention the type of information you consulted on how to use Autopilot and FSD Beta (website of Tesla (www.tesla.com), car dealer / sales point, online communities and forums, YouTube videos, newspapers and magazines, friends, family, colleagues, driver manual)
Q4	Please describe your experience with using Autopilot and FSD Beta and the benefits and risks associated with using it. Please explain your answer
Q5	Have your expectations of using Autopilot and FSD Beta been fulfilled? Why / why not?
Q6	Why do you use Autopilot and FSD Beta?
Q7	Did you ever stop using Autopilot and FSD Beta (for prolonged periods of time)?
Next, we would like to explore your perceptions regarding four general statements about the operation of Autopilot and FSD Beta	
Q8	The current Autopilot does make driving autonomous. Is that correct? (1 = Yes, 2 = No, 3 = I don't know)
Q9	There are no safety issues with Autopilot. Is that correct? (1 = Yes, 2 = No, 3 = I don't know)
Q10	Autopilot is a hands-on feature. Is that correct? (1 = Yes, 2 = No, 3 = I don't know)
Q11	Tesla FSD Beta is safer than a human. Is that correct? (1 = Yes, 2 = No, 3 = I don't know)
With the next section, we would like to explore your perceptions of safety while using Autopilot and FSD Beta	
Q12	Do you feel safe when Autopilot and FSD Beta is active? Why / why not?
Q13	What / how do you feel when you feel safe / unsafe? Please explain
Q14	What is it about Autopilot and FSD Beta that is safe / unsafe? Please explain.
Q15	Now please remember the situation / s in which you typically feel unsafe when Autopilot and FSD Beta is active and describe these situations.
Q16	What can Autopilot and FSD Beta do to support your safety in Autopilot and FSD Beta? Please explain
Q17	Does feeling safe / feeling unsafe impact how you use Autopilot and FSD Beta on your next drives / in the future? Please explain.
Q18	Has your perceived safety changed over time? If so, how?
With the next section, we would like to explore your trust in Autopilot and FSD Beta.	
Q19	How would you position your level of trust in Autopilot and FSD Beta. (1 = I don't trust it at all, 2 = I don't trust it, 3 = I neither don't trust it at all nor trust it a lot, 4 = I trust it, 5 = I trust it a lot)

Continued on next page

Table 2 – Continued from previous page

Question number	Question
Q20	What can Autopilot and FSD Beta do to support your trust in Autopilot and FSD Beta?
Q21	Does your trust / distrust in Autopilot and FSD Beta impact how you use Autopilot and FSD Beta on your next drives / in the future? Please explain.
Q22	Has your trust changed over time? If so, how?
Q23	When you do compare yourself with other drivers, Autopilot, and FSD Beta, do you think you are ... (1 = A much worse driver, 2 = A worse driver, 3 = Not a better nor a worse driver, 4 = A better driver, 5 = A much better driver) (De Craen, 2010)
With the next section, we would like to explore how you typically use Autopilot and FSD Beta.	
Q24	How do you typically place your hands on the steering wheel when Autopilot and FSD Beta is active? Please select the image that serves as the best representation of your placement of your hands on the steering wheel when Autopilot / FSD Beta is active and explain your answer.
Q25	Do you typically keep your hands on the steering wheel at all times?
Q26	Are you typically fully attentive and alert at all times?
Q27	How often do you typically engage in other secondary activities while Autopilot and FSD Beta is active? (Never, rarely, occasionally, frequently, always; monitoring the road ahead, talking to fellow travelers, observing the landscape, using the phone for music selection, using the phone for navigation, using the phone for calls, eating and drinking, using the phone for texting, watching videos / TV shows, sleeping)
Q28	Do you disengage Autopilot and FSD Beta? Why / why not?
Q29	Does Autopilot and FSD Beta disengage? When / in which situations?
Q30	How do you typically place your eyes when Autopilot and FSD Beta is active?
Q31	Do you typically keep your eyes on the road at all times?
Q32	Do you typically monitor the vehicle and its surroundings at all times?
Q33	How do you typically place your feet when Autopilot and FSD Beta is active?
Q34	Do you typically stay prepared to take corrective actions at all times?
Q35	Has your use of Autopilot (in terms of how you placed your hands on the steering wheel, eyes on the road, and feet) changed over time? If so, how?

C Appendix for Chapter 8: Objective Assessment of Meaningful Human Control

C.1 Survey Questions for MHC Concepts

Post experiment questions	Subjective perception
D1: Were there any situations where you felt you *had sufficient control* over the automated vehicle operation? Please describe.	Sufficient control
D2: Were there any situations where you felt you *did not have sufficient control* over the automated vehicle operation? Please describe.	Sufficient control
D3: Were there any situations where you felt that the automated vehicle *understood your intention*? Please describe.	AV understood me
D4: Were there any situations where you felt that the automated vehicle *did not understand your intention*? Please describe.	AV understood me
D5: Were there any situations where you felt that you *understood the behaviour of the automated vehicle*? Please describe.	I understood AV
D6: Were there any situations where you felt that you *did not understand the behaviour of the automated vehicle*? Please describe.	I understood AV
D7: Were there any situations where you felt that the automated vehicle was *working together with you*? Please describe.	Working together
D8: Were there any situations where you felt that the automated vehicle was *not working together with you*? Please describe.	Working together
D9: Were there any situations where you *felt responsible* for the driving task when using the automated vehicle? Please describe.	Responsible
D10: Were there any situations where you *did not feel responsible* for the driving task when using the automated vehicle? Please describe.	Resposible
D11: Were there any situations where you felt *safe* when using the automated vehicle? Please describe.	Safe
D12: Were there any situations where you felt *unsafe* when using the automated vehicle? Please describe.	Safe
D13: Were there any situations where you *trusted* the automated vehicle? Please describe.	Trust
D14: Were there any situations where you *did not trust* the automated vehicle? Please describe.	Trust

Table 3: Descriptive questions after all the trials

C.2 Calculating Behavioural Metrics

This appendix provides the full preprocessing steps, equations, and references for the behavioural metrics summarised in Section 7.3.3. Each subsection below corresponds to a specific

metric.

Preprocessing

All telemetry signals were preprocessed before analysis. Specifically, we applied a centered rolling mean with a 20-sample window, corresponding to 0.4 s at the simulator's 50 Hz sampling rate. This smoothing reduced random measurement noise while preserving the overall signal dynamics. For notational simplicity, we denote signal as continuous function of time (e.g., $\tau_H(t)$, although in practice they were recorded as discrete samples at 50 Hz. From these smoothed signals, we then derived the behavioural metrics described below and explain their relevance for assessing MHC. Additionally, we applied a torque threshold, retaining only human and automation torque values with magnitude greater than 0.2 Nm to reduce noise. Throughout the remainder of this section, 'torque' refers to these thresholded values.

Reaction time

Reaction time (RT) was defined as the delay between the automation initiating an evasive manoeuvre toward a potential collision with a motorcyclist and the participant's takeover response. Formally,

$$RT = t_H - t_A,$$

where t_A denotes the automation initiation time and t_H denotes the human response time. Because the automation acts first, reaction time is defined such that $t_H > t_A$.

Reaction time was computed for both Haptic Shared Control (HSC) and Traded Control (TC). The automation initiation time t_A was defined identically for both control modes, whereas the human response time t_H was operationalised differently depending on the control strategy.

The automation initiation time (t_A) was identified as the moment when the automation torque began a rapid change indicating the onset of an evasive manoeuvre. Specifically, t_A was defined as the first time at which the gradient of the automation torque exceeded a threshold of 0.18 Nm/s:

$$t_A = \min\left(t \mid \frac{d\tau_A(t)}{dt} > 0.18 \text{ and } t < t_{\text{peak}}\right),$$

where $\tau_A(t)$ denotes the automation torque and t_{peak} is the time of the first local maximum of the torque gradient following manoeuvre onset. The first peak was selected because it corresponds to the moment when the automation action becomes most perceptually salient to the driver.

The definition of human response time (t_H) depended on the control mode:

- **Haptic Shared Control (HSC):** t_H was defined as the first moment after t_A at which conflict between human and automation was detected,

$$t_H^{HSC} = \min\left(t \mid C(t) > 0 \text{ and } t > t_A\right),$$

where $C(t)$ denotes the interaction conflict signal between human and automation inputs, as defined in C.3.

- **Traded Control (TC):** t_H was defined as the first moment after t_A at which automation authority $A(t)$ dropped below a predefined threshold θ, indicating transfer of control to the human driver. The automation authority signal $A(t)$ is defined in C.3:

$$t_H^{TC} = \min\left(t \mid A(t) < \theta \text{ and } t > t_A\right).$$

According to **?**, sufficient control can be interpreted as the driver's ability to intervene and redirect the vehicle toward safety without being constrained by the automation. Shorter reaction times therefore indicate faster driver responsiveness once the automation initiates a potentially unsafe trajectory.

Figure 1 and 2 illustrate the extraction procedure. Panel (a) shows detection of t_A from the automation torque gradient, while panel (b) illustrates identification of t_H based on conflict onset (HSC) or authority reduction (TC).

C.3 Conflict

Conflict ($C(t)$) in steering torques was defined as instances where human and the automation applied torques in opposite directions. Formally:

$$C(t) = -\tau_H(t) \cdot \tau_A(t)$$

where $\tau_H(t)$ is the human-applied torque and $\tau_A(t)$ is the automation-applied torque. Positive conflict values indicate opposing torques.

We included this metric because conflict can signal a lack of cooperation between human and automation. According to Cavalcante Siebert et al. (2023), shared representation, where human and AI agents maintain mutually compatible task, facilitates cooperation. High conflict, by contrast, suggests misalignment and poorer coordination.

To ensure consistency with the controller design used in the experiment, we used the same torque threshold that TC used to trigger authority handover to the human. This event-based operationalisation aligns with the prior work that interprets increased opposing steering torques as a proxy for human–automation conflict (Boink et al., 2014). Figure 3 illustrates both the temporal evolution of the conflict signal and its spatial distribution along the driven trajectory, highlighting where human corrective actions emerged during interaction with the automation.

Maximum steering torque

Maximum steering torque was defined as the highest torque value applied by the human driver ($\tau_H(t)$) during a trial. This metric reflects how much physical effort the driver exerted to override or resist the automation, defined as the following definition:

$$\text{Maximum steering torque} = \max(\tau_H(t))$$

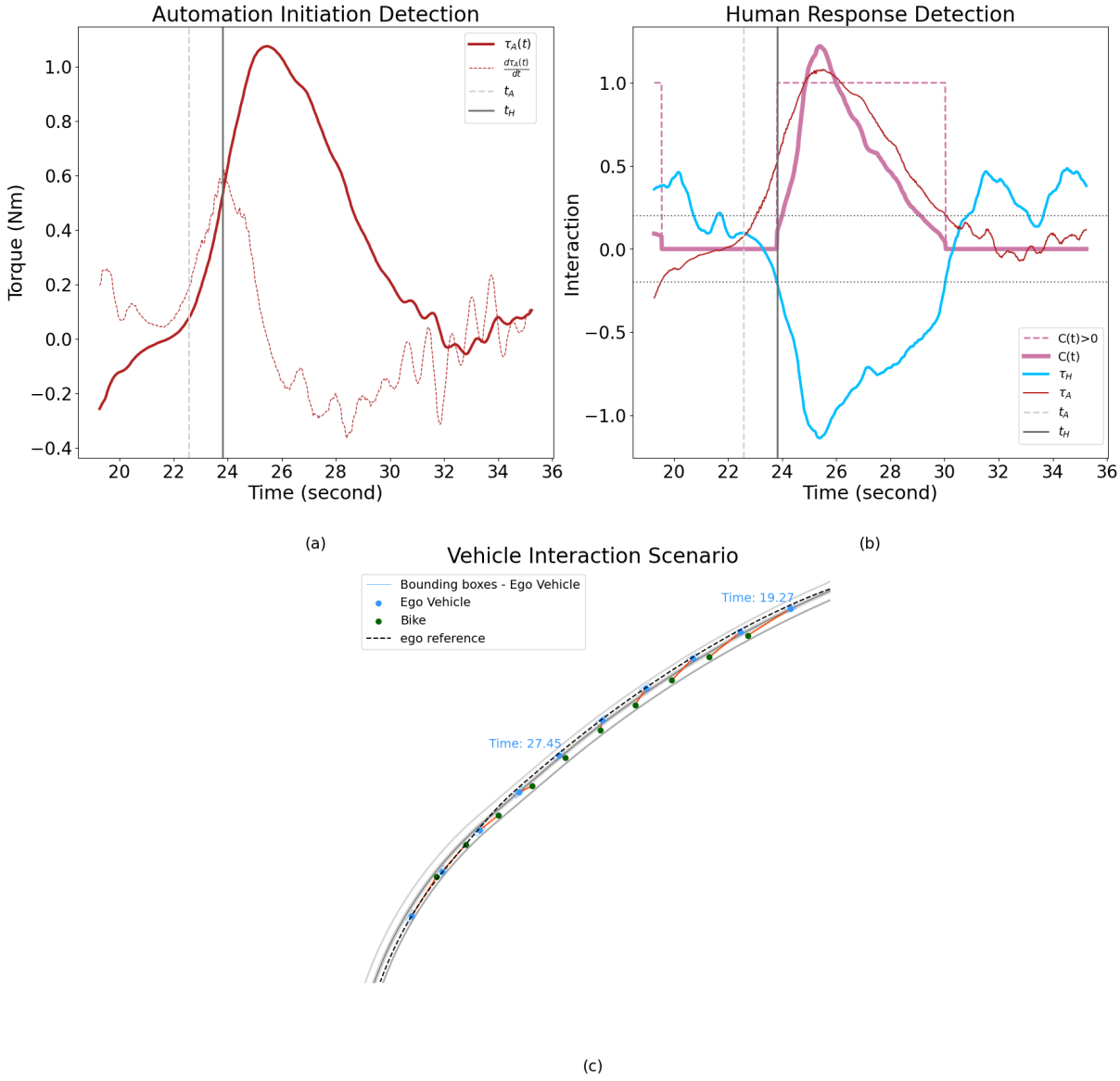


Figure 1: Illustration of reaction-time calculation for Haptic Shared Control. (a) Detection of automation initiation time t_A from the automation torque gradient $\frac{d\tau_a(t)}{dt}$. (b) Identification of human response time t_H based on conflict onset. (c) Illustration of ego vehicle and bike position over time. The pair of positions of the ego and bike for the same time stamp is shown by the orange line between them. The black dashed line is the ego vehicle reference trajectory. The time stamps denote two representative sampling instants: an earlier reference moment and the instant at which the distance between the ego vehicle and the bike is minimal.

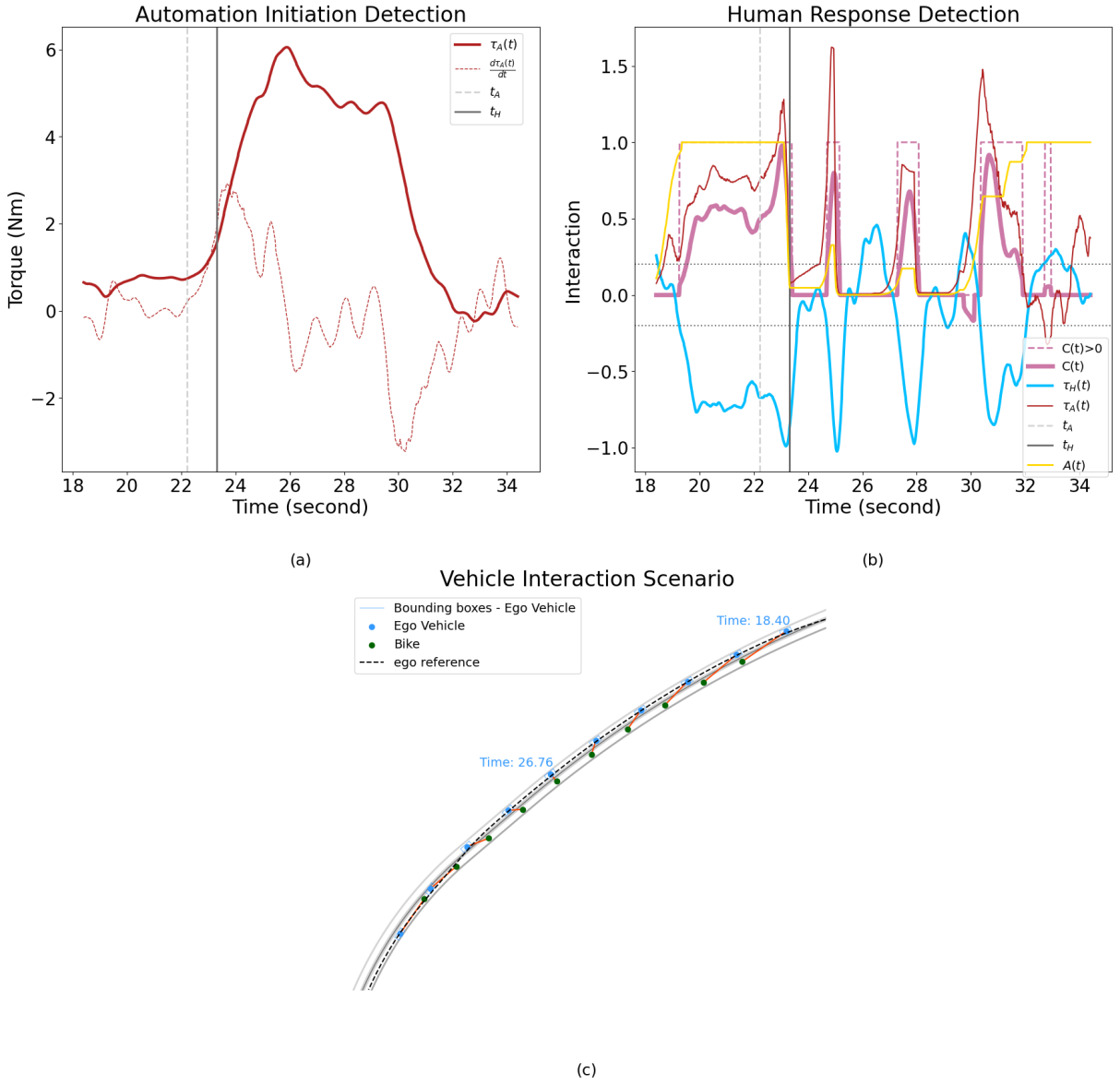


Figure 2: Illustration of reaction-time calculation for Traded Control. (a) Detection of automation initiation time t_A from the automation torque gradient $\frac{d\tau_a(t)}{dt}$. (b) Identification of human response time t_H based authority reduction. (c) Illustration of ego vehicle and bike position over time.

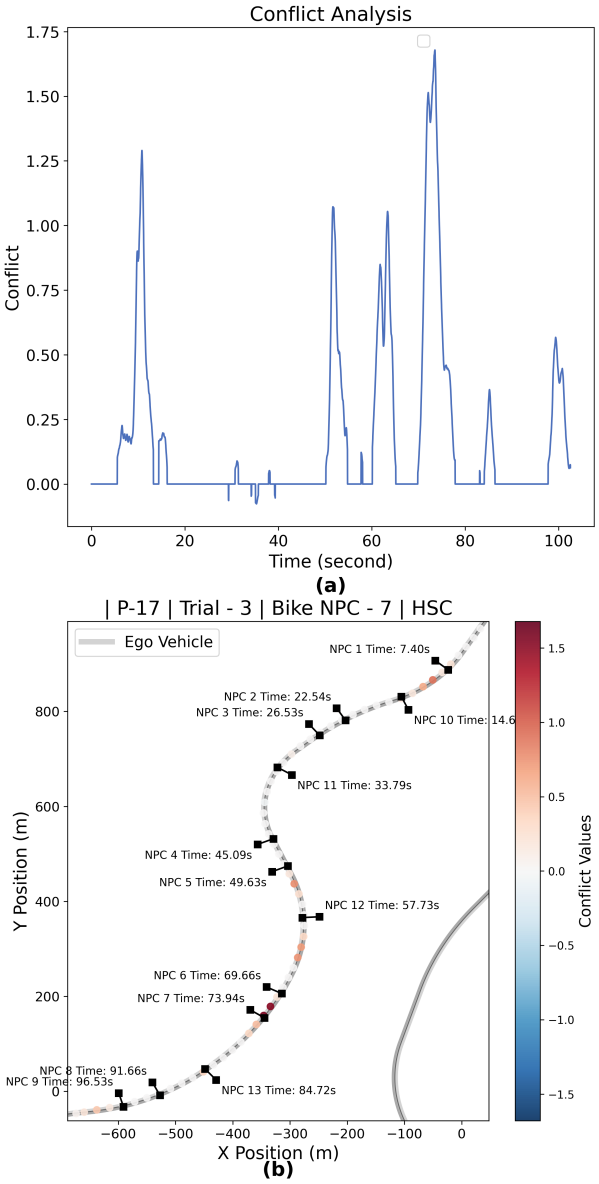


Figure 3: Illustration of human–automation conflict during a representative trial under Haptic Shared Control (HSC). (a) Temporal evolution of the conflict signal $C(t)$ over the entire driving trial. Peaks indicate periods of opposing steering torques between the human driver and the automation. (b) Spatial distribution of conflict values along the ego vehicle trajectory. Conflict magnitude is colour-coded, where darker red indicates higher conflict intensity. Black markers denote traffic participants. Markers positioned on the left side of the trajectory indicate vehicles travelling in the same direction as the ego vehicle, whereas markers on the right indicate opposing traffic.

.

We included this measure because high torque demands may indicate poor cooperation: if drivers must exert stronger force against the system, they are likely to feel less "in control" and less as though they are "working together" with the automation. Accordingly, we hypothesised negative associations between maximum steering torque and perceptions of both sufficient control and cooperation.

Prior research supports this interpretation. For example, Ercan et al. (2018) found that drivers applied maximum torque when resisting lane-keeping or lane-departure assistance, situations directly related to perceived control and cooperation.

Steering reversal rate

Steering reversal rate was calculated as the number of steering reversals per second as defined by Markkula & Engström (2006), where a steering reversal is defined as the portion of the interactions where the rate of change of the steering angle is zero, and the difference in steering angle at the two extremes is greater than 1°.

We included steering reversal rate as a behavioural metric as it is a widely accepted metric for control effort while driving (Markkula & Engström, 2006; Mars et al., 2014).

Take overs

We defined two ways to calculate the number of takeovers because haptic shared control (HSC) and traded control (TC) differ in nature. For HSC, a takeover occurs when the conflict value changes from zero to positive. For TC, a takeover occurs when authority $A(t)$ falls below a set threshold after previously being above it. After defining these triggers, we counted the number of takeover events per trial.

In our framework, authority represents how much control the automation currently has, ranging from 0 (driver full control) to 1 (automation full control). Authority is calculated from the driver's steering torque: if the driver applies strong torque, authority decreases, showing that the driver is taking over. If the driver applies little torque for some time, authority increases, showing that the automation regains control. In this way, authority provides a continuous measure of who is in control at any given moment.

Formally, authority is expressed as:

$$A(t) = 1 - \left(1 + e^{-c_1\,(x(t)-c_2)}\right)^{-1},$$

where $x(t)$ is an internal signal that is updated based on the driver's torque input. When the human torque τ_h exceeds a threshold, x decreases sharply (leading authority to drop). When τ_H stays well below the threshold, x increases gradually (leading authority to rise). The constants c_1 and c_2 control the steepness and midpoint of the transition.

We used the number of takeovers as an indicator of how drivers perceive the AV to collaborate with them. Prior work by Hwang et al. (2025) indicates that when drivers are engaged in

more complex secondary tasks, their takeover performance deteriorates, with slower responses and lower success rates. Extending this logic, drivers who perceive the automation as competent or "understanding" may become less prepared to intervene, resulting in fewer takeovers. This complacency can be seen as an extreme version of collaboration, in which a greater portion of the driving task is left to the automation. Taken together, this supports the use of the number of takeovers to capture drivers' perception of collaboration with the AV.

Trajectory deviation

Trajectory deviation (TD) was defined as root mean square error (RMSE) between the actual trajectory of the ego vehicle and automation's trajectory during a trial:

$$TD = \sqrt{\frac{1}{N}\Sigma_{i=1}^{N}[(x_{actual}(i) - x_{ref}(j))^2 - (y_{actual}(i) - y_{ref}(j))^2]}$$

where $x_{ref}(j)$ and $y_{ref}(j)$ are the reference trajectory points closest in position to the actual $x_{actual}(i)$ and $y_{actual}(i)$.

Prior work (Griesche et al., 2016) shows that most drivers prefer to drive in their own style or in a style similar to it. Extending this finding, we expected that if drivers felt responsible for operational control but found the automation's trajectory deviated from their preferred style, they would be more likely to take over. This takeover would increase trajectory deviation but, at the same time, might strengthen their perceived sense of responsibility. Figure 4 illustrates the spatial deviation between the actual and reference trajectories.

Jerk

Jerk (J) was defined as a measure of driving smoothness, obtained by differentiating the vehicle's acceleration and calculating its root mean square (RMS) across one trial::

$$J = \sqrt{\frac{1}{N}\Sigma_{i=1}^{N}(j_{x,i}^2 + j_{y,i}^2)}$$

where $j_{x,i}$ and $j_{y,i}$ denote the jerk samples in the longitudinal and lateral directions, respectively. High jerk values indicate abrupt changes in acceleration, which can reflect uncomfortable driving or difficulties in controlling the automation. Accordingly, higher jerk were interpreted as reduced perceptions of sufficient control. Figure 5 illustrates the spatial distribution of jerk values along the driven trajectory.

Overtaking time

Overtaking time was defined as the total duration the vehicle spent in the overtaking lane during a trial. We calculated the signed lateral offset d_i of the vehicle from the road centerline at each sample by taking the Euclidean distance to the nearest centerline point and assigning a sign using the centerline's local normal:

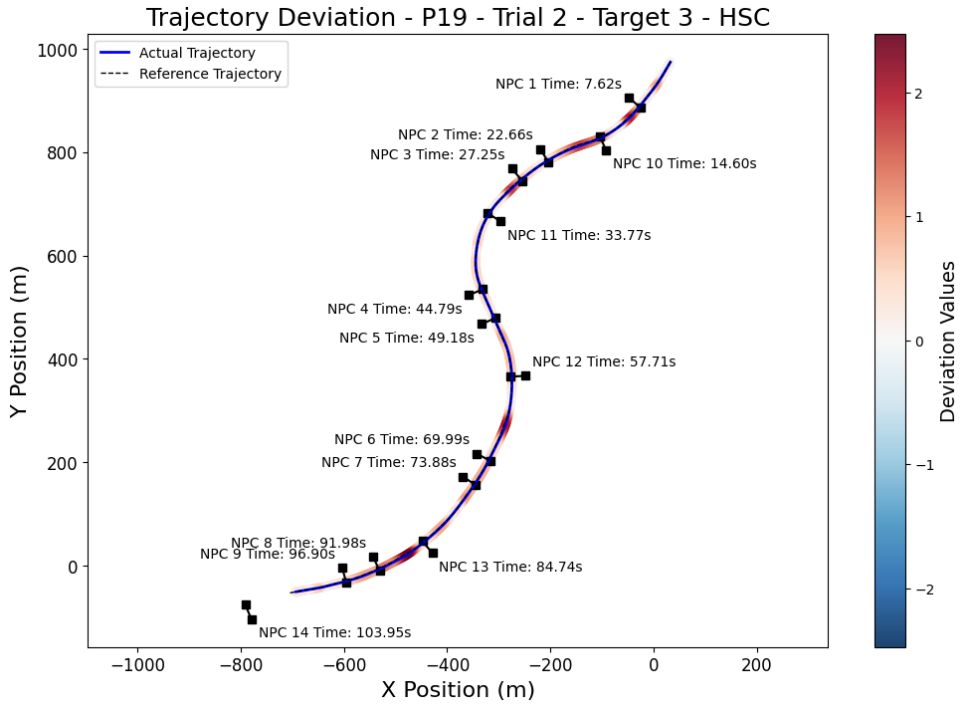


Figure 4: Illustration of trajectory deviation calculation. The solid blue line shows the actual ego-vehicle trajectory, while the dashed black line denotes the automation reference trajectory. Deviation values are computed as the pointwise Euclidean distance between the actual and reference trajectories and are colour-coded along the driven path.

$$d_i = ||\mathbf{p}_i - \mathbf{c}_{\text{nearest}}|| \times \text{sign}((\mathbf{p}_i - \mathbf{c}_{\text{nearest}}) \cdot \mathbf{n}_{\text{nearest}})$$

where $\mathbf{p}_i$vehicle position (x_i, y_i)at time i, $\mathbf{c}_{\text{nearest}}$ is nearest point on road centerline, $\mathbf{n}_{\text{nearest}}$ perpendicular unit vector at nearest road point, and $\cdot$ is a dot product. Positive values of d_i indicate that the vehicle is in the overtaking lane. Overtaking time was then calculated by summing the durations of all contiguous intervals where $d_i > 0$:

$$T_{\text{over}} = \sum_{k=1}^{K} \left(t_{i_x^{(k)}} - t_{i_e^{(k)}} \right),$$

where $i_x^{(k)}$ and $i_e^{(k)}$ are the enter and exit indices of overtaking intervals and t_i are the recorded timestamps. For each time point i, we calculated the vehicle's lateral distance d_i from the road centerline by finding the nearest centerline point and computing the signed perpendicular distance. The sign was determined using the dot product with the road's perpendicular vector, where positive values indicate the overtaking lane.

Although not included in the hypotheses, we included overtaking time post-hoc as a cue for collaboration. In both HSC and TC, automation followed its own overtaking trajectory and typically merged back after passing. Longer overtaking times therefore suggest that drivers either resisted the automation's intent to return to the driving lane (HSC) or withheld authority to prolong the maneuver (TC), indicating reduced collaboration. Shorter overtaking times are more consistent with alignment with the automation's planned trajectory, supporting the perception

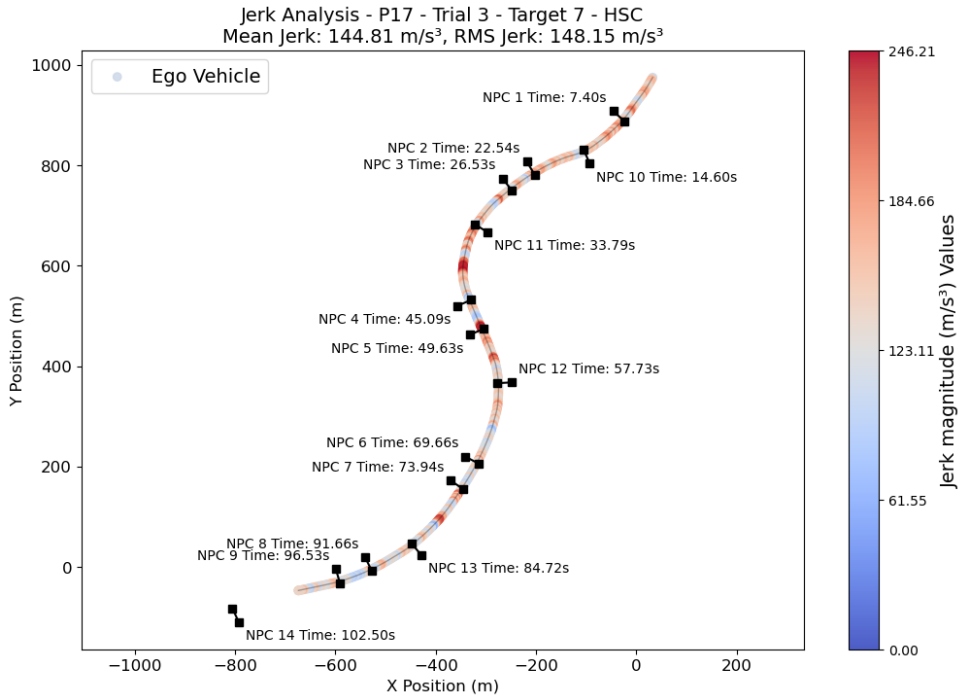


Figure 5: Illustration of jerk calculation. The ego-vehicle trajectory is colour-coded according to jerk magnitude, computed from the time derivative of vehicle acceleration. Colours represent pointwise jerk values along the trajectory, while the reported mean and RMS jerk values are calculated over all samples in the trial.

of working together. Figure 6 illustrates the identification of overtaking intervals based on the signed lateral offset and the aggregation of these intervals used to compute overtaking time.

Time to Collision

Time to collision of the ego vehicle to the target vehicle was calculated using the algorithm proposed by Jiao (2023) to calculate two-dimensional TTC. The minimum TTC value per trial was used for the analysis, as it would represent the most critical part of the interaction. In many cases, the velocities of the vehicles were never directed towards a collision and resulted in infinite values of the minimum TTC. Before we ran the statistical analysis we winsorised the data by replacing infinite values with the 0.99 percentile values of the non-infinite values.

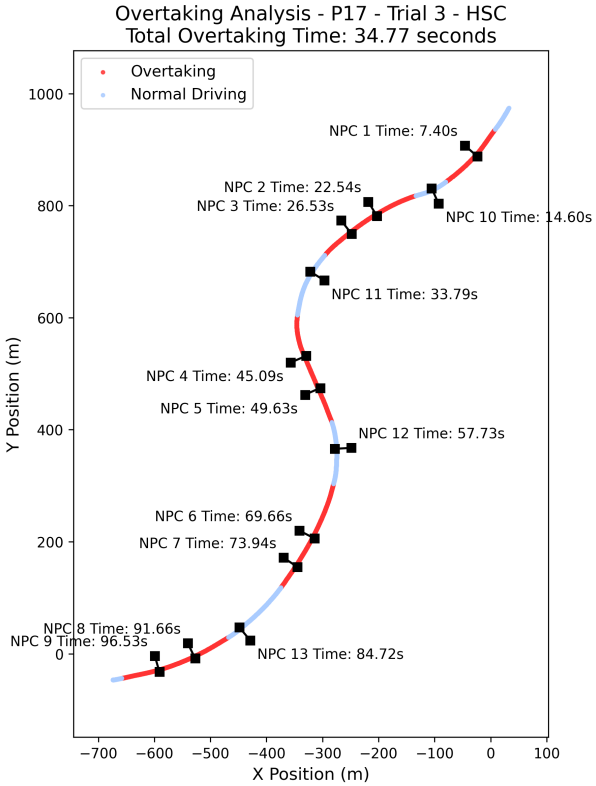


Figure 6: Illustration of overtaking time calculation. The ego-vehicle trajectory is colour-coded according to driving state, where red segments indicate periods in which the vehicle occupies the overtaking lane ($d_i > 0$) and blue segments indicate normal driving. Overtaking is determined using the signed lateral offset from the road centerline, computed at each time step. Total overtaking time is obtained by summing the durations of all contiguous overtaking intervals.

Bibliography

Aasvik, O., M. Hagenzieker, P. Ulleberg (2025) I trust norway – investigating acceptance of shared autonomous shuttles using open and closed questions in short-form street interviews, *Transportation Research Interdisciplinary Perspectives*, 31, p. 101414.

Abbink, D. A., T. Carlson, M. Mulder, J. C. De Winter, F. Aminravan, T. L. Gibo, E. R. Boer (2018) A topology of shared control systems—finding common ground in diversity, *IEEE Transactions on Human-Machine Systems*, 48(5), pp. 509–525.

Abbink, D. A., M. Mulder, E. R. Boer (2012) Haptic shared control: smoothly shifting control authority?, *Cognition, Technology & Work*, 14, pp. 19–28.

Agriesti, S., F. Brevi, P. Gandini, G. Marchionni, R. Parmar, M. Ponti, L. Studer (2020) Impact of driverless vehicles on urban environment and future mobility, *Transportation Research Procedia*, 49, pp. 44–59.

Aksjonov, A., V. Kyrki (2021) Rule-based decision-making system for autonomous vehicles at intersections with mixed traffic environment, in: *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, IEEE, pp. 660–666.

Asaro, P. (2012) On banning autonomous weapon systems: Human rights, automation, and the dehumanization of lethal decision-making, *International Review of the Red Cross*, 94(886), p. 687–709.

Awad, E., S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, I. Rahwan (2018) The moral machine experiment, *Nature*, 563(7729), pp. 59–64.

Awad, E., S. Levine, M. Kleiman-Weiner, S. Dsouza, J. B. Tenenbaum, A. Shariff, J.-F. Bonnefon, I. Rahwan (2020) Drivers are blamed more than their automated cars when both make mistakes, *Nature human behaviour*, 4(2), pp. 134–143.

Bainbridge, L. (1983) Ironies of automation, in: *Analysis, design and evaluation of man–machine systems*, Elsevier, pp. 129–135.

Banks, G. C., H. M. Woznyj, R. S. Wesslen, R. L. Ross (2018a) A review of best practice recommendations for text analysis in r (and a user-friendly app), *Journal of Business and Psychology*, 33, pp. 445–459.

Banks, V. A., A. Eriksson, J. O'Donoghue, N. A. Stanton (2018b) Is partially automated driving a bad idea? Observations from an on-road study, *Applied ergonomics*, 68, pp. 138–145.

Bays, P. M., J. R. Flanagan, D. M. Wolpert (2006) Attenuation of Self-Generated Tactile Sensations Is Predictive, not Postdictive, *PLoS Biology*, 4(2), p. e28.

Beckers, N., L. C. Siebert, M. Bruijnes, C. Jonker, D. Abbink (2022) Drivers of partially automated vehicles are blamed for crashes that they cannot reasonably avoid, *Scientific Reports*, 12(1), p. 16193.

Beckers, N., O. Siebinga, J. Giltay, A. van der Kraan (2023) Joan: A framework for human-automated vehicle interaction experiments in a virtual reality driving simulator, *Journal of Open Source Software*, 8(82), p. 4250.

Bellet, T., S. Laurent, J.-C. Bornard, I. Hoang, B. Richard (2022) Interaction between pedestrians and automated vehicles: Perceived safety of yielding behaviors and benefits of an external human–machine interface for elderly people, *Frontiers in psychology*, 13, p. 1021656.

Berberian, B., J.-C. Sarrazin, P. Le Blaye, P. Haggard (2012a) Automation technology and sense of control: a window on human agency, *PloS one*, 7(3), p. e34075.

Berberian, B., J.-C. Sarrazin, P. Le Blaye, P. Haggard (2012b) Automation Technology and Sense of Control: A Window on Human Agency, *PLoS ONE*, 7(3), p. e34075.

Bergmann, L. T. (2022) Ethical issues in automated driving—opportunities, dangers, and obligations, *User experience design in the era of automated driving*, pp. 99–121.

Beringhoff, F., J. Greenyer, C. Roesener, M. Tichy (2022) Thirty-one challenges in testing automated vehicles: Interviews with experts from industry and research, in: *2022 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, pp. 360–366.

Bin-Nun, A. Y., P. Derler, N. Mehdipour, R. D. Tebbens (2022) How should autonomous vehicles drive? policy, methodological, and social considerations for designing a driver, *Humanities and social sciences communications*, 9(1), pp. 1–13.

Boink, R., M. M. Van Paassen, M. Mulder, D. A. Abbink (2014) Understanding and reducing conflicts between driver and haptic shared control, in: *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, pp. 1510–1515.

Bonnefon, J.-F., D. Černy, J. Danaher, N. Devillier, V. Johansson, T. Kovacikova, M. Martens, M. Mladenovic, P. Palade, N. Reed, et al. (2020) Ethics of connected and automated vehicles: Recommendations on road safety, privacy, fairness, explainability and responsibility, *Placeholder Journal*.

Bonnefon, J.-F., A. Shariff, I. Rahwan (2016) The social dilemma of autonomous vehicles, *Science*, 352(6293), pp. 1573–1576.

Bonnefon, J.-F., A. Shariff, I. Rahwan (2019) The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars [point of view], *Proceedings of the IEEE*, 107(3), pp. 502–504.

Borowsky, A., D. Shinar, T. Oron-Gilad (2010) Age, skill, and hazard perception in driving, *Accident analysis & prevention*, 42(4), pp. 1240–1249.

Braun, V., V. Clarke (2006) Using thematic analysis in psychology, *Qualitative research in psychology*, 3(2), pp. 77–101.

Burcu, A. (2000) A comparison of two data collecting methods: interviews and questionnaires, *Hacettepe Univ J Educ*, 18, pp. 1–10.

Cabrall, C. D., A. Eriksson, F. Dreger, R. Happee, J. de Winter (2019) How to keep drivers engaged while supervising driving automation? a literature survey and categorisation of six solution areas, *Theoretical issues in ergonomics science*, 20(3), pp. 332–365.

Calvert, S., D. D. Heikoop, G. Mecacci, B. Van Arem (2020a) A human centric framework for the analysis of automated driving systems based on meaningful human control, *TheoreTical issues in ergonomics science*, 21(4), pp. 478–506.

Calvert, S., S. Johnsen, A. George (2024) Designing automated vehicle and traffic systems towards meaningful human control, in: *Research handbook on meaningful human control of artificial intelligence systems*, Edward Elgar Publishing, pp. 162–187.

Calvert, S., G. Mecacci (2020) A conceptual control system description of cooperative and automated driving in mixed urban traffic with meaningful human control for design and evaluation, *IEEE Open Journal of Intelligent Transportation Systems*, 1, pp. 147–158.

Calvert, S., G. Mecacci, D. D. Heikoop, R. Janssen (2021) How to ensure control of cooperative vehicle and truck platoons using meaningful human control, *European Journal of Transport and Infrastructure Research*, 21(2), pp. 95–119.

Calvert, S., B. van Arem, D. D. Heikoop, M. Hagenzieker, G. Mecacci, F. S. de Sio (2020b) Gaps in the control of automated vehicles on roads, *IEEE intelligent transportation systems magazine*, 13(4), pp. 146–153.

Calvert, S. C. (2025) Principles and framework for the operationalisation of meaningful human control over autonomous systems, *Science and Engineering Ethics*, 31(5), p. 27.

Campbell, J. L., C. Quincy, J. Osserman, O. K. Pedersen (2013) Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement, *Sociological methods & research*, 42(3), pp. 294–320.

Cavalcante Siebert, L., M. L. Lupetti, E. Aizenberg, N. Beckers, A. Zgonnikov, H. Veluwenkamp, D. Abbink, E. Giaccardi, G.-J. Houben, C. M. Jonker, et al. (2023) Meaningful human control: actionable properties for ai system development, *AI and Ethics*, 3(1), pp. 241–255.

Cecchini, D., M. Pflanzer, V. Dubljević (2024) Aligning artificial intelligence with moral intuitions: An intuitionist approach to the alignment problem, *AI and Ethics*, pp. 1–11.

Charmaz, K. (2014) *Constructing grounded theory*, SAGE publications Ltd.

Chen, K., Z. Li, P. Liu, V. L. Knoop, Y. Han, Y. Jiao (2024) Evaluating the safety and efficiency impacts of forced lane change with negative gaps based on empirical vehicle trajectories, *Accident Analysis & Prevention*, 203, p. 107622.

Choi, J. K., Y. G. Ji (2015) Investigating the importance of trust on adopting an autonomous vehicle, *International Journal of Human-Computer Interaction*, 31(10), pp. 692–702.

Chu, Y., P. Liu (2023) Automation complacency on the road, *Ergonomics*, 66(11), pp. 1730–1749.

Cicchino, J. B. (2017) Effectiveness of forward collision warning and autonomous emergency braking systems in reducing front-to-rear crash rates, *Accident Analysis & Prevention*, 99, pp. 142–152.

Cohen, J. (1960) A coefficient of agreement for nominal scales, *Educational and psychological measurement*, 20(1), pp. 37–46.

Connected Automated Driving (2024) Taxonomy of automated driving systems, URL https://www.connectedautomateddriving.eu/methodology/taxonomy/, accessed: 2025-02-02.

Cornelio, P., P. Haggard, K. Hornbaek, O. Georgiou, J. Bergström, S. Subramanian, M. Obrist (2022) The sense of agency in emerging technologies for human–computer integration: A review, *Frontiers in Neuroscience*, 16, p. 949138.

Cummings, M. M. (2025) Identifying ai hazards and responsibility gaps, *IEEE Access*.

Cummings, M. M., B. Wheeler, J. Kliem (2024) A root cause analysis of a self-driving car dragging a pedestrian, *Computer*, 57(11), pp. 31–40.

Cunningham, A. G., E. Galceran, R. M. Eustice, E. Olson (2015) Mpdm: Multipolicy decision-making in dynamic, uncertain environments for autonomous driving, in: *2015 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, pp. 1670–1677.

D'Amato, A., S. Dancel, J. Pilutti, L. Tellis, E. Frascaroli, J. Gerdes (2022) Exceptional driving principles for autonomous vehicles, *JL & Mobility*, p. 1.

Dang, R., J. Wang, S. E. Li, K. Li (2015) Coordinated adaptive cruise control system with lane-change assistance, *IEEE Transactions on Intelligent Transportation Systems*, 16(5), pp. 2373–2383.

de Sio, F. S., G. Mecacci, S. Calvert, D. Heikoop, M. Hagenzieker, B. van Arem (2023) Realising meaningful human control over automated driving systems: a multidisciplinary approach, *Minds and machines*, 33(4), pp. 587–611.

de Winter, J. C., S. M. Petermeijer, D. A. Abbink (2023) Shared control versus traded control in driving: a debate around automation pitfalls, *Ergonomics*, 66(10), pp. 1494–1520.

Detjen, H., M. Salini, J. Kronenberger, S. Geisler, S. Schneegass (2021) Towards transparent behavior of automated vehicles: Design and evaluation of hud concepts to support system predictability through motion intent communication, in: *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*, pp. 1–12.

Dixit, V. V., S. Chand, D. J. Nair (2016) Autonomous vehicles: disengagements, accidents and reaction times, *PLoS one*, 11(12), p. e0168054.

Docherty, B. L. (2015) Mind the gap: The lack of accountability for killer robots, https://www.hrw.org/report/2015/04/09/mind-gap/lack-accountability-killer-robots, accessed: 21 September 2025.

Dosovitskiy, A., G. Ros, F. Codevilla, A. Lopez, V. Koltun (2017) Carla: An open urban driving simulator, in: *Conference on robot learning*, PMLR, pp. 1–16.

Dreger, F. A., J. C. de Winter, R. Happee (2020) How do drivers merge heavy goods vehicles onto freeways? a semi-structured interview unveiling needs for communication and support, *Cognition, Technology & Work*, 22(4), pp. 825–842.

Dubljević, V., S. Douglas, J. Milojevich, N. Ajmeri, W. A. Bauer, G. List, M. P. Singh (2023) Moral and social ramifications of autonomous vehicles: a qualitative study of the perceptions of professional drivers, *Behaviour & Information Technology*, 42(9), pp. 1271–1278.

Endsley, M. R. (2017) Autonomous driving systems: A preliminary naturalistic study of the tesla model s, *Journal of Cognitive Engineering and Decision Making*, 11(3), pp. 225–238.

Ercan, Z., A. Carvalho, H. E. Tseng, M. Gökaşan, F. Borrelli (2018) A predictive control framework for torque-based steering assistance to improve safety in highway driving, *Vehicle system dynamics*, 56(5), pp. 810–831.

European Union (2018) Regulation (EU) 2018/858 of the European Parliament and of the Council of 30 May 2018 on the approval and market surveillance of motor vehicles and their trailers, URL https://eur-lex.europa.eu/eli/reg/2018/858/oj, official Journal of the European Union, L 151, 14.6.2018, p. 1–218.

Fagnant, D. J., K. Kockelman (2015) Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations, *Transportation Research Part A: Policy and Practice*, 77, pp. 167–181.

Feinstein, A. R., D. V. Cicchetti (1990) High agreement but low kappa: I. the problems of two paradoxes, *Journal of clinical epidemiology*, 43(6), pp. 543–549.

Felzmann, H., E. Fosch-Villaronga, C. Lutz, A. Tamò-Larrieux (2020) Towards transparency by design for artificial intelligence, *Science and engineering ethics*, 26(6), pp. 3333–3361.

Flemisch, F., M. Heesen, T. Hesse, J. Kelsch, A. Schieben, J. Beller (2012) Towards a dynamic balance between humans and automation: Authority, ability, responsibility and control in shared and cooperative control situations, *Cognition, Technology & Work*, 14(1), pp. 3–18.

Fraade-Blanar, L., M. S. Blumenthal, J. M. Anderson, N. Kalra (2018) *Measuring Automated Vehicle Safety: Forging a Framework*, RAND Corporation, Santa Monica, Calif., library of Congress Cataloging-in-Publication Data available.

FSDEvolution (2025) Tesla following a cyclist, URL https://www.youtube.com/shorts/ATFAxNikP_8, accessed: 2025-03-01.

Geisslinger, M., F. Poszler, J. Betz, C. Lütge, M. Lienkamp (2021) Autonomous driving ethics: From trolley problem to ethics of risk, *Philosophy & Technology*, 34(4), pp. 1033–1055.

Geisslinger, M., F. Poszler, M. Lienkamp (2023) An ethical trajectory planning algorithm for autonomous vehicles, *Nature Machine Intelligence*, 5(2), pp. 137–144.

George, A., L. C. Siebert, D. Abbink, A. Zgonnikov (2023) Feasible action-space reduction as a metric of causal responsibility in multi-agent spatial interactions, *arXiv preprint arXiv:2305.15003*.

Graneheim, U., B. Lundman (2004) Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness, *Nurse Education Today*, 24(2), pp. 105–112.

Griesche, S., E. Nicolay, D. Assmann, M. Dotzauer, D. Käthner (2016) Should my car drive as i do? what kind of driving style do drivers prefer for the design of automated driving functions, in: *Braunschweiger symposium*, vol. 10, pp. 185–204.

Gros, C., L. Kester, M. Martens, P. Werkhoven (2025a) Modelling societal preferences for automated vehicle behaviour with ethical goal functions, *Frontiers in Artificial Intelligence*, 8, p. 1676225.

Gros, C., P. Werkhoven, L. Kester, M. Martens (2025b) A methodology for ethical decision-making in automated vehicles, *AI & SOCIETY*, pp. 1–12.

Gwet, K. L. (2008) Computing inter-rater reliability and its variance in the presence of high agreement, *British Journal of Mathematical and Statistical Psychology*, 61(1), pp. 29–48.

Habibullah, K. M., H.-M. Heyn, G. Gay, J. Horkoff, E. Knauss, M. Borg, A. Knauss, H. Sivencrona, P. J. Li (2024) Requirements and software engineering for automotive perception systems: an interview study, *Requirements Engineering*, 29(1), pp. 25–48.

Hagemeister, C., L. Bertram (2024) Reported pushy driving against cyclists in germany, *Journal of safety research*, 88, pp. 395–405.

Hansson, S. O., M.-Å. Belin, B. Lundgren (2021) Self-driving vehicles—an ethical overview, *Philosophy & Technology*, 34(4), pp. 1383–1408.

He, X., J. Stapel, M. Wang, R. Happee (2022) Modelling perceived risk and trust in driving automation reacting to merging and braking vehicles, *Transportation research part F: traffic psychology and behaviour*, 86, pp. 178–195.

Heikoop, D. D., M. Hagenzieker, G. Mecacci, S. Calvert, F. Santoni De Sio, B. van Arem (2019) Human behaviour with automated driving systems: a quantitative framework for meaningful human control, *Theoretical issues in ergonomics science*, 20(6), pp. 711–730.

Henschke, A. (2020) Trust and resilient autonomous driving systems, *Ethics and Information Technology*, 22(1), pp. 81–92.

Hickman, L., S. Thapa, L. Tay, M. Cao, P. Srinivasan (2022) Text preprocessing for text mining in organizational research: Review and recommendations, *Organizational Research Methods*, 25(1), pp. 114–146.

Hilgarter, K., P. Granig (2020) Public perception of autonomous vehicles: A qualitative study based on interviews after riding an autonomous shuttle, *Transportation research part F: traffic psychology and behaviour*, 72, pp. 226–243.

Hill, C. E., B. J. Thompson, E. N. Williams (2005) Consensual qualitative research: An update, *Journal of Counseling Psychology*, 52(2), pp. 196–205.

Himmelreich, J. (2018) Never mind the trolley: The ethics of autonomous vehicles in mundane situations, *Ethical Theory and Moral Practice*, 21(3), pp. 669–684.

Homayounirad, A., E. Liscia, T. Wang, C. M. Jonker, L. C. Siebert (2025) Will annotators disagree? identifying subjectivity in value-laden arguments, in: *Findings of the Association for Computational Linguistics: EMNLP 2025*, forthcoming.

Horowitz, M. C., P. Scharre (2015) Meaningful human control in weapon systems: A primer, Working paper, Center for a New American Security.

Hsieh, H.-F., S. E. Shannon (2005) Three approaches to qualitative content analysis, *Qualitative Health Research*, 15(9), pp. 1277–1288.

Huang, Z., J. Wu, C. Lv (2021) Driving behavior modeling using naturalistic human driving data with inverse reinforcement learning, *IEEE transactions on intelligent transportation systems*, 23(8), pp. 10239–10251.

Hwang, J., W. Choi, J. Lee, W. Kim, J. Rhim, A. Kim (2025) A dataset on takeover during distracted l2 automated driving, *Scientific Data*, 12(1), p. 539.

International Organization for Standardization (2019) *ISO/PAS 21448: Road vehicles - Safety of the intended functionality*, Geneva, Switzerland, pre-standards document.

Jiang, L., Y. Xie, N. G. Evans (2023) A simulation study of cooperative and autonomous vehicles (cav) considering courtesy, ethics, and fairness, *Plos one*, 18(5), p. e0283649.

Jiao, Y. (2023) A fast calculation of two-dimensional Time-to-Collision, URL https://github.com/Yiru-Jiao/Two-Dimensional-Time-To-Collision.

Keeling, G. (2020) Why trolley problems matter for the ethics of automated vehicles, *Science and engineering ethics*, 26(1), pp. 293–307.

Kim, S., X. He, R. van Egmond, R. Happee (2024) Designing user interfaces for partially automated vehicles: Effects of information and modality on trust and acceptance, *Transportation research part F: traffic psychology and behaviour*, 103, pp. 404–419.

Kirchmair, L., N. Paulo (2023) Taking ethics seriously in av trajectory planning algorithms, *Nature Machine Intelligence*, 5(8), pp. 814–815.

Koglbauer, I., J. Holzinger, A. Eichberger, C. Lex (2018) Autonomous emergency braking systems adapted to snowy road conditions improve drivers' perceived safety and trust, *Traffic injury prevention*, 19(3), pp. 332–337.

Koopman, P., W. H. Widen (2023) A reasonable driver standard for automated vehicle safety, in: *International Conference on Computer Safety, Reliability, and Security*, Springer, pp. 355–361.

Kraus, J., D. Scholz, D. Stiegemeier, M. Baumann (2020) The more you know: trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency, *Human factors*, 62(5), pp. 718–736.

Krippendorff, K. (2018) *Content Analysis: An Introduction to Its Methodology*, 4 ed., SAGE Publications.

Kusano, K. D., J. M. Scanlon, Y.-H. Chen, T. L. McMurry, T. Gode, T. Victor (2025) Comparison of waymo rider-only crash rates by crash type to human benchmarks at 56.7 million miles, *Traffic Injury Prevention*, pp. 1–13.

Kwik, J. (2022) A practicable operationalisation of meaningful human control, *Laws*, 11(3), p. 43.

Lee, J. D., K. A. See (2004) Trust in automation: Designing for appropriate reliance, *Human factors*, 46(1), pp. 50–80.

Lee, S. C., C. Nadri, H. Sanghavi, M. Jeon (2020) Exploring user needs and design requirements in fully automated vehicles, in: *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, pp. 1–9.

Lee, Y.-C. (2010) Measuring drivers' frustration in a driving simulator, in: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 54, Sage Publications Sage CA: Los Angeles, CA, pp. 1531–1535.

Lee, Y. M., R. Madigan, T. Louw, E. Lehtonen, N. Merat (2023) Does users' experience and evaluation of level 3 automated driving functions predict willingness to use: Results from an on-road study, *Transportation research part F: traffic psychology and behaviour*, 99, pp. 473–484.

Leenes, R., F. Lucivero (2014) Laws on robots, laws by robots, laws in robots: Regulating robot behaviour by design, *Law, Innovation and Technology*, 6(2), pp. 193–220.

Li, J., X. Zhao, M.-J. Cho, W. Ju, B. F. Malle (2016) From trolley to autonomous vehicle: Perceptions of responsibility and moral norms in traffic accidents with self-driving cars, Tech. rep., SAE Technical Paper.

Li, M., H. Cao, X. Song, Y. Huang, J. Wang, Z. Huang (2018) Shared control driver assistance system based on driving intention and situation assessment, *IEEE Transactions on Industrial Informatics*, 14(11), pp. 4982–4994.

Lin, P. (2016) Why ethics matters for autonomous cars, *Autonomous driving: Technical, legal and social aspects*, pp. 69–85.

Liu, M., Y. Wan, F. L. Lewis, S. Nageshrao, D. Filev (2022) A three-level game-theoretic decision-making framework for autonomous vehicles, *IEEE Transactions on Intelligent Transportation Systems*, 23(11), pp. 20298–20308.

Liu, P., J. Liu (2021) Selfish or utilitarian automated vehicles? deontological evaluation and public acceptance, *International Journal of Human–Computer Interaction*, 37(13), pp. 1231–1242.

Ljubi, K., A. Groznik (2023) Role played by social factors and privacy concerns in autonomous vehicle adoption, *Transport policy*, 132, pp. 1–15.

Longhurst, R., L. Johnston (2023) 10 semi-structured interviews and focus groups, *Key methods in geography*, 168.

Lu, H., M. Zhu, C. Lu, S. Feng, X. Wang, Y. Wang, H. Yang (2025) Empowering safer socially sensitive autonomous vehicles using human-plausible cognitive encoding, *Proceedings of the National Academy of Sciences*, 122(21).

Ma, J., X. Feng (2024) Analysing the effects of scenario-based explanations on automated vehicle hmis from objective and subjective perspectives, *Sustainability*, 16(1), p. 63.

Ma, Z., Y. Zhang (2021) Drivers trust, acceptance, and takeover behaviors in fully automated vehicles: Effects of automated driving styles and driver's driving styles, *Accident Analysis & Prevention*, 159.

Markkula, G., J. Engström (2006) A steering wheel reversal rate metric for assessing effects of visual and cognitive secondary task load, in: *Proceedings of the 13th ITS World Congress*, Leeds.

Mars, F., M. Deroo, J.-M. Hoc (2014) Analysis of human-machine cooperation when driving with different degrees of haptic shared control, *IEEE transactions on haptics*, 7(3), pp. 324–333.

Martinho, A., N. Herber, M. Kroesen, C. Chorus (2021) Ethical issues in focus by the autonomous vehicles industry, *Transport reviews*, 41(5), pp. 556–577.

Matin, A., H. Dia (2022) Impacts of connected and automated vehicles on road safety and efficiency: A systematic literature review, *IEEE Transactions on Intelligent Transportation Systems*, 24(3), pp. 2705–2736.

Matthias, A. (2004) The responsibility gap: Ascribing responsibility for the actions of learning automata, *Ethics and information technology*, 6(3), pp. 175–183.

Mecacci, G., S. C. Calvert, F. Santoni de Sio (2023) Human–machine coordination in mixed traffic as a problem of meaningful human control, *AI & society*, 38(3), pp. 1151–1166.

Mecacci, G., F. Santoni de Sio (2020) Meaningful human control as reason-responsiveness: the case of dual-mode vehicles, *Ethics and Information Technology*, 22, pp. 103–115.

Meder, B., N. Fleischhut, N.-C. Krumnau, M. R. Waldmann (2019) How should autonomous cars drive? a preference for defaults in moral judgments under risk and uncertainty, *Risk analysis*, 39(2), pp. 295–314.

Mehta, D., G. Ferrer, E. Olson (2016) Autonomous navigation in dynamic social environments using multi-policy decision making, in: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, pp. 1190–1197.

Merat, N., H. A. Jamson, F. Lai, O. Carsten (2014) Human factors of highly automated driving: results from the easy and citymobil projects, in: *Road vehicle automation*, Springer, pp. 113–125.

Michon, J. A. (1985) A critical view of driver behavior models: what do we know, what should we do?, in: *Human behavior and traffic safety*, Springer, pp. 485–524.

Milford, S. R., B. Z. Malgir, B. S. Elger, D. M. Shaw (2025) "All things equal": Ethical principles governing why autonomous vehicle experts change or retain their opinions in trolley problems—a qualitative study, *Frontiers in Robotics and AI*, 12.

Millar, J., P. Lin, K. Abney, G. Bekey (2017) Ethics settings for autonomous vehicles, *Robot ethics*, 2, pp. 20–34.

Miller, E. E., L. N. Boyle (2019) Behavioral adaptations to lane keeping systems: Effects of exposure and withdrawal, *Human factors*, 61(1), pp. 152–164.

Molnar, L. J., L. H. Ryan, A. K. Pradhan, D. W. Eby, R. M. S. Louis, J. S. Zakrajsek (2018) Understanding trust and acceptance of automated vehicles: An exploratory simulator study of transfer of control between automated and manual driving, *Transportation research part F: traffic psychology and behaviour*, 58, pp. 319–328.

Montoro, L., S. A. Useche, F. Alonso, I. Lijarcio, P. Bosó-Seguí, A. Martí-Belda (2019) Perceived safety and attributed value as predictors of the intention to use autonomous vehicles: A national study with spanish drivers, *Safety Science*, 120, pp. 865–876.

Moore, J. W. (2016) What is the sense of agency and why does it matter?, *Frontiers in psychology*, 7, p. 1272.

Moretto, G., E. Walsh, P. Haggard (2011) Experience of agency and sense of responsibility, *Consciousness and cognition*, 20(4), pp. 1847–1854.

Mulder, M., D. A. Abbink, E. R. Boer (2012) Sharing control with haptics: Seamless driver support from manual to automatic control, *Human factors*, 54(5), pp. 786–798.

National Transportation Safety Board (2017) Collision between a car operating with automated vehicle control systems and a tractor-semitrailer truck near williston, Tech. rep., Accident Report NTSB/HAR-17/02 PB2017-102600.

National Transportation Safety Board (2018) Collision between a sport utility vehicle operating with partial driving automation and a crash attenuator, Tech. rep., Accident Report NTSB/HAR-20/01 PB2020-100112.

Naveteur, J., P. Delhomme, C. Terrier (2013) Impatience and time pressure: Subjective reactions of drivers in situations forcing them to stop their car in the road, *Transportation Research Part F: Traffic Psychology and Behaviour*, 18, pp. 58–71.

NLTK Project (2024) Natural Language Toolkit, https://www.nltk.org, [Online; accessed: 19-October-2023].

Nordhoff, S. (2024) A conceptual framework for automation disengagements, *Scientific Reports*, 14, p. 8654.

Nordhoff, S., J. De Winter (2023) Why do drivers and automation disengage the automation? results from a study among tesla users, *arXiv preprint arXiv:2309.10440*.

Nordhoff, S., M. Hagenzieker (2024) "I will raise my hand and say 'I over-trust autopilot'. I use it too liberally"–Drivers' reflections on their use of partial driving automation, trust, and perceived safety, *Transportation Research Part F: Traffic Psychology and Behaviour*, 107, pp. 1105–1124.

Nordhoff, S., J. D. Lee, S. Calvert, S. Berge, M. Hagenzieker, R. Happee (2023) (mis-) use of standard autopilot and full self-driving (fsd) beta: results from interviews with users of tesla's fsd beta, *Frontiers in psychology*, 14, p. 1101520.

Nordhoff, S., J. Stapel, X. He, A. Gentner, R. Happee (2021) Perceived safety and trust in sae level 2 partially automated cars: Results from an online questionnaire, *Plos one*, 16(12), p. e0260953.

Nyholm, S., J. Smids (2016) The ethics of accident-algorithms for self-driving cars: An applied trolley problem?, *Ethical theory and moral practice*, 19(5), pp. 1275–1289.

Nyholm, S., J. Smids (2020) Automated cars meet human drivers: responsible human-robot coordination and the ethics of mixed traffic, *Ethics and Information Technology*, 22, pp. 335–344.

Olleja, P., G. Markkula, J. Bärgman (2025) Validation of human benchmark models for automated driving system approval: How competent and careful are they really?, *Accident Analysis & Prevention*, 213, p. 107922.

Oskina, M., H. Farah, P. Morsink, R. Happee, B. van Arem (2023) Safety assessment of the interaction between an automated vehicle and a cyclist: a controlled field test, *Transportation research record*, 2677(2), pp. 1138–1149.

Papadimitriou, E., H. Farah, G. van de Kaa, F. Santoni de Sio, M. Hagenzieker, P. van Gelder (2022) Towards common ethical and safe 'behaviour' standards for automated vehicles, *Accident Analysis & Prevention*, 174, p. 106724.

Paula, D., M. Bauder, C. Pfeilschifter, F. Petermeier, T. Kubjatko, K. Böhm, A. Riener, H.-G. Schweiger (2023) Impact of partially automated driving functions on forensic accident reconstruction: A simulator study on driver reaction behavior in the event of a malfunctioning system behavior.

Payre, W., J. Cestac, P. Delhomme (2016) Fully automated driving: Impact of trust and practice on manual control recovery, *Human factors*, 58(2), pp. 229–241.

Peng, C., S. Horn, R. Madigan, C. Marberger, J. D. Lee, J. Krems, M. Beggiato, R. Romano, C. Wei, E. Wooldridge, R. Happee, M. Hagenzieker, N. Merat (2024) Conceptualising user comfort in automated driving: Findings from an expert group workshop, *Transportation Research Interdisciplinary Perspectives*, 24, p. 101070.

Peng, C., N. Merat, R. Romano, F. Hajiseyedjavadi, E. Paschalidis, C. Wei, V. Radhakrishnan, A. Solernou, D. Forster, E. Boer (2022) Drivers' evaluation of different automated driving styles: Is it both comfortable and natural?, *Human factors*.

Quinn Emanuel Urquhart & Sullivan LLP (2024) Report to the boards of directors of cruise llc, gm cruise holdings llc, and general motors holdings llc regarding the october 2, 2023 accident in san francisco, Tech. rep., Quinn Emanuel Urquhart & Sullivan LLP, privileged and Confidential Internal Report.

Rahmani, S., S. Calvert, B. van Arem (2025) Decentralized modeling of vehicular maneuvers and interactions at urban junctions, *https://arxiv.org/abs/2507.21547*.

Rahmani, S., J. Neumann, L. E. Suryana, C. Theunisse, S. Calvert, B. Van Arem (2023) A bi-level real-time microsimulation framework for modeling two-dimensional vehicular maneuvers at intersections, in: *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, pp. 4221–4226.

Rhim, H. J., J. M. Urban (2021) A deeper look at moral dilemmas in autonomous driving: The integrative ethical decision-making framework, *Frontiers in Robotics and AI*, 8.

Road Safety Authority (2018) Examining the international research evidence in relation to minimum passing distances for cyclists, Tech. rep., Road Safety Authority (RSA) of Ireland, pre-legislative scrutiny report, Road Safety Research and Driver Education.

Robbins, S. (2023) The many meanings of meaningful human control, *AI and Ethics*, pp. 1–12.

Robins-Early, N. (2024) Driver in fatal Texas crash was using Ford's auto driving system, officials say, https://www.theguardian.com/business/2024/apr/11/ford-mustang-mach-e-automated-blue-cruise-system-fatal-texas-crash, [Online; accessed 24-April-2024].

Saber, E. M., S.-C. Kostidis, I. Politis (2024) Ethical dilemmas in autonomous driving: Philosophical, social, and public policy implications, *Springer*, pp. 7–20.

SAE International (2021) Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems, Standard J3016_202104.

Santoni de Sio, F., G. Mecacci (2021) Four responsibility gaps with artificial intelligence: Why they matter and how to address them, *Philosophy & technology*, 34(4), pp. 1057–1084.

Santoni de Sio, F., J. Van den Hoven (2018) Meaningful human control over autonomous systems: A philosophical account, *Frontiers in Robotics and AI*, 5, p. 15.

Schreier, M. (2012) *Qualitative content analysis in practice*, Sage.

Schwarting, W., J. Alonso-Mora, D. Rus (2018) Planning and decision-making for autonomous vehicles, *Annual Review of Control, Robotics, and Autonomous Systems*, 1(1), pp. 187–210.

Schwarting, W., A. Pierson, J. Alonso-Mora, S. Karaman, D. Rus (2019) Social behavior for autonomous vehicles, *Proceedings of the National Academy of Sciences*, 116(50), pp. 24972–24978.

SensoDrive (2025) Sensodrive force feedback, URL https://www.sensodrive.de/products/force-feedback-products.php, accessed: 2025-11-08.

Shea-Blymyer, C., H. Abbas (2021) Algorithmic ethics: Formalization and verification of autonomous vehicle obligations, *ACM Transactions on Cyber-Physical Systems (TCPS)*, 5(4), pp. 1–25.

Steen, M., J. van Diggelen, T. Timan, N. van der Stap (2023) Meaningful human control of drones: exploring human–machine teaming, informed by four different ethical perspectives, *AI and Ethics*, 3(1), pp. 281–293.

Sun, L., W. Zhan, M. Tomizuka, A. D. Dragan (2018) Courteous autonomous cars, in: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, pp. 663–670.

Suryana, L. E., S. Calvert, A. Zgonnikov, B. van Arem (2025a) Principles and reasons behind automated vehicle decisions in ethically ambiguous everyday scenarios, *arXiv preprint arXiv:2507.13837*.

Suryana, L. E., S. Nordhoff, S. Calvert, A. Zgonnikov, B. van Arem (2025b) Meaningful human control of partially automated driving systems: Insights from interviews with tesla users, *Transportation Research Part F: Traffic Psychology and Behaviour*, 113, pp. 213–236.

Suryana, L. E., S. Nordhoff, S. C. Calvert, A. Zgonnikov, B. Van Arem (2024) A meaningful human control perspective on user perception of partially automated driving systems: a case study of tesla users, in: *2024 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, pp. 409–416.

Suryana, L. E., S. Rahmani, S. Calvert, A. Zgonnikov, B. van Arem (2025c) A framework for human-reason-aligned trajectory evaluation in automated vehicles, *arXiv preprint arXiv:2507.23324*.

Suryana, L. E., S. Rahmani, S. C. Calvert, A. Zgonnikov, B. Van Arem (2025d) A human reasons-based supervision framework for ethical decision-making in automated vehicles, in: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, to appear.

Swain, R., V. Truelove, A. Rakotonirainy, S.-A. Kaye (2023) A comparison of the views of experts and the public on automated vehicles technologies and societal implications, *Technology in Society*, 74, p. 102288.

Tabone, W., J. De Winter, C. Ackermann, J. Bärgman, M. Baumann, S. Deb, C. Emmenegger, A. Habibovic, M. Hagenzieker, P. A. Hancock, et al. (2021) Vulnerable road users and the coming wave of automated vehicles: Expert perspectives, *Transportation research interdisciplinary perspectives*, 9, p. 100293.

Teng, S., X. Hu, P. Deng, B. Li, Y. Li, Y. Ai, D. Yang, L. Li, Z. Xuanyuan, F. Zhu, et al. (2023) Motion planning for autonomous driving: The state of the art and future perspectives, *IEEE Transactions on Intelligent Vehicles*, 8(6), pp. 3692–3711.

Tesla (2024a) Autopilot and Full Self-Driving Capability, https://www.tesla.com/support/autopilot, [Online; accessed 19-October-2024].

Tesla (2024b) Model 3 Owner's Manual, https://www.tesla.com/ownersmanual/model3, [Online; accessed: 19-August-2024].

Tesla (2025) Update vehicle firmware to disable fsd beta “rolling stop” functionality, accessed: 2025-04-13.

Thornton, S. M., F. E. Lewis, V. Zhang, M. J. Kochenderfer, J. C. Gerdes (2018) Value sensitive design for autonomous vehicle motion planning, in: *2018 IEEE intelligent vehicles symposium (IV)*, IEEE, pp. 1157–1162.

Thornton, S. M., S. Pan, S. M. Erlien, J. C. Gerdes (2016) Incorporating ethical considerations into automated vehicle control, *IEEE Transactions on Intelligent Transportation Systems*, 18(6), pp. 1429–1439.

Tversky, A., D. Kahneman (1992) Advances in prospect theory: Cumulative representation of uncertainty, *Journal of Risk and uncertainty*, 5, pp. 297–323.

Umbrello, S., R. V. Yampolskiy (2022) Designing ai for explainability and verifiability: a value sensitive design approach to avoid artificial stupidity in autonomous vehicles, *International Journal of Social Robotics*, 14(2), pp. 313–322.

UNECE (2023) UN Regulation No. 157: Uniform provisions concerning the approval of vehicles with regard to Automated Lane Keeping Systems, URL https://unece.org/transport/documents/2023/03/standards/un-regulation-no-157-amend4, accessed: 2024-12-07.

Vaismoradi, M., H. Turunen, T. Bondas (2013) Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study, *Nursing & health sciences*, 15(3), pp. 398–405.

Veluwenkamp, H. (2022) Reasons for meaningful human control, *Ethics and Information Technology*, 24(4), p. 51.

Verhagen, R. S., M. A. Neerincx, M. L. Tielman (2024) Meaningful human control and variable autonomy in human-robot teams for firefighting, *Frontiers in Robotics and AI*, 11, p. 1323980.

von Stülpnagel, R., J. Lucas (2020) Crash risk and subjective risk perception during urban cycling: Evidence for congruent and incongruent sources, *Accident Analysis & Prevention*, 142, p. 105584.

Wang, H., Y. Huang, A. Khajepour, D. Cao, C. Lv (2020) Ethical decision-making platform in autonomous vehicles with lexicographic optimization based model predictive controller, *IEEE transactions on vehicular technology*, 69(8), pp. 8164–8175.

Wang, J., C. Yu, S. E. Li, L. Wang (2015) A forward collision warning algorithm with adaptation to driver behaviors, *IEEE Transactions on Intelligent Transportation Systems*, 17(4), pp. 1157–1167.

Wang, Z., R. Zheng, T. Kaizuka, K. Shimono, K. Nakano (2017) The effect of a haptic guidance steering system on fatigue-related driver behavior, *IEEE Transactions on Human-Machine Systems*, 47(5), pp. 741–748.

Watanabe, K., Y. Zhou (2022) Theory-driven analysis of large corpora: Semisupervised topic classification of the un speeches, *Social Science Computer Review*, 40(2), pp. 346–366.

Wen, W., H. Imamizu (2022) The sense of agency in perception, behaviour and human–machine interactions, *Nature Reviews Psychology*, 1(4), pp. 211–222.

Wenke, D., P. Haggard (2009) How voluntary actions modulate time perception, *Experimental Brain Research*, 196(3), pp. 311–318.

Williams, G., A. Aldrich, E. A. Theodorou (2017) Model predictive path integral control: From theory to parallel computation, *Journal of Guidance, Control, and Dynamics*, 40(2), pp. 344–357.

Wörle, J., B. Metz (2023) Misuse or abuse of automation? exploring drivers' intentions to nap during automated driving, *Transportation research part F: traffic psychology and behaviour*, 99, pp. 460–472.

Wynne, R. A., V. Beanland, P. M. Salmon (2019) Systematic review of driving simulator validation studies, *Safety science*, 117, pp. 138–151.

Xu, W., J. Wei, J. M. Dolan, H. Zhao, H. Zha (2012) A real-time motion planner with trajectory optimization for autonomous vehicles, in: *2012 IEEE international conference on robotics and automation*, IEEE, pp. 2061–2067.

Xu, Z., K. Zhang, H. Min, Z. Wang, X. Zhao, P. Liu (2018) What drives people to accept automated vehicles? findings from a field experiment, *Transportation research part C: emerging technologies*, 95, pp. 320–334.

Young, M. S., N. A. Stanton (2002) Malleable attentional resources theory: a new explanation for the effects of mental underload on performance, *Human factors*, 44(3), pp. 365–375.

Yuan, K., Y. Huang, S. Yang, Z. Zhou, Y. Wang, D. Cao, H. Chen (2024) Evolutionary decision-making and planning for autonomous driving based on safe and rational exploration and exploitation, *Engineering*, 33, pp. 108–120.

Zhang, L., Y. Dong, H. Farah, B. van Arem (2023) Social-aware planning and control for automated vehicles based on driving risk field and model predictive contouring control: Driving through roundabouts as a case study, in: *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, pp. 3297–3304.

Zhang, Q., C. D. Wallbridge, D. M. Jones, P. L. Morgan (2024) Public perception of autonomous vehicle capability determines judgment of blame and trust in road traffic accidents, *Transportation research part A: policy and practice*, 179, p. 103887.

Zhang, T., D. Tao, X. Qu, X. Zhang, R. Lin, W. Zhang (2019) The roles of initial trust and perceived risk in public's acceptance of automated vehicles, *Transportation research part C: emerging technologies*, 98, pp. 207–220.

Acknowledgements

Completing this PhD has been one of the most challenging and rewarding experiences of my life. The past four years have witnessed my transformation from a highly technical person into someone who has learned to look beyond technical solutions, appreciating that the most meaningful answers often lie at the intersection of disciplines. All milestones achieved in this PhD would not have been possible without the support, guidance, and encouragement of many people, to whom I am deeply grateful. Some may say that I was lucky to be surrounded by such people, but what I truly feel is that I was blessed.

My deepest gratitude goes first to my supervisory team, whose guidance and dedication have been instrumental in bringing this thesis to completion. **Bart**, I see you as a true professional role model in academia. Despite your demanding schedule, you were always present at every meeting and consistently provided timely feedback. Your openness created a space where I felt comfortable sharing anything. I also want to mention the Dutch conversations you always initiated at the start of our meetings; they quietly but meaningfully helped me grow more confident in speaking the language. **Simeon**, thank you for opening the door to this journey. I still remember the day I received your reply to my email about the PhD opportunity. You were warm and welcoming from the very first call. You introduced me to the topic of meaningful human control, gave me the freedom to shape it, and showed me that you believed in my ability to carry it forward. That trust means more to me than words can express. It planted in me a confidence that I could take on challenges beyond what I had imagined for myself and an openness to embrace opportunities I would not have dared to consider before. Thank you also for connecting me with collaborators whose contributions turned out to be central to the completion of this thesis. **Arkady**, although you joined my supervisory team a little later, your impact has been profound. Your feedback consistently pushed my work one step further than I thought possible, and your availability always made me feel supported. From you, I also learned the value of critical thinking and giving feedback that is truly thorough and precise. This is a skill I will carry well beyond this PhD. Thank you as well for the opportunity to collaborate within the Center for Meaningful Human Control, a topic I find as exciting as it is meaningful.

I would also like to extend my sincere gratitude to the doctoral committee for their time and willingness to evaluate my work. **Filippo**, it is a true honor to have been evaluated by one of the main authors of the reference paper on meaningful human control. In our several interactions, you confirmed some of my ideas, and I greatly appreciate this, as it has meaningfully shaped my current work. **Marieke**, thank you for your detailed feedback. It opened my eyes to an entire line of research addressing similar issues, and your insights helped shape my work into something far better than it would have been otherwise. **Javier**, thank you for your helpful evaluation. I am also deeply grateful for the permission to attend your Planning and Decision-Making class. The knowledge I gained there proved so valuable that it eventually developed

into two chapters of this thesis. **Nynke**, thank you for agreeing to be part of my committee. I thoroughly enjoyed the invitation to visit Groningen and explore how meaningful human control relates to the legal field. It was an eye-opening experience.

I am deeply grateful to the **Indonesia Endowment Fund for Education (LPDP)** for the trust and financial support that made this journey possible. The opportunity to pursue a PhD at one of the world's leading technical universities is a privilege I do not take lightly, and I hope the knowledge and experience gained here will one day find its way back to Indonesia in a right way.

To my colleagues in 4.02 (DiTTLab), both those still in the lab and those who have moved on, you have all been far more than colleagues. You have been wonderful friends who accompanied me through these four years. **Xue**, I truly enjoy the conversations that happen every time I come to the lab. From you I have learned how to stay productive in research while also embracing the enjoyable things life has to offer. **Yiru**, you are warm, friendly, and always have something interesting to talk about. I have also learned a lot from you about academic excellence. Beyond the scientific contribution, your research work is always visually compelling. **Kexin**, you are a warm and fun person to talk to. **Srinath**, thank you for your friendliness and your willingness to strike up a conversation. I especially cherish the meal at the Indian restaurant in Utrecht and your visit to my place. **Saeed**, many people say you are the friendliest person in the department, and I have no reason to disagree. I have learned so much from you about warmth, reliability, and an incredible work ethic. Who would have thought that taking a Planning & Decision Making class together would lead to three published papers? **Samir**, it has been a genuine pleasure to first meet you as a master's student and then become colleagues in the same room. Congratulations to Iraq for qualifying for the 2026 World Cup! It is always a joy to catch up and hear about your research. You are truly an out-of-the-box thinker. I also want to mention **Khaqan**, **Ali**, **Zili**, **Yiyun**, **Mingze**, **Xiaolin**, **Robin**, **Yiyun**, **Chaopeng**, **Wouter**, **Felix**, **Guopeng**, **Zara**, **Tin**, **Tamim**, **Jonah** and **Mariko**. Thank you all for making DiTTLab such a warm and stimulating place to work.

To my collaborators, I would like to express my sincere appreciation. **Sina**, I have greatly enjoyed working with you. From you I learned qualitative research methods and how to work efficiently and write methodology sections with clarity and precision. I sincerely hope you secure a tenured position very soon. **Ashwin**, for nearly four years we have worked together on something that felt like it would take forever, and it is now almost there (almost!). I have no doubt that outside of research we make great companions, though inside research we may be a different story. You know exactly what I mean, and I say that with a smile. **Holger**, thank you for your enthusiasm and quick response to the new research idea I am developing around VLM reasoning. I deeply appreciate your openness, and I believe this will be an important step in my career going forward. I hope it leads to something truly meaningful. **Luciano**, thank you for being willing to meet and talk about career directions. Your generosity in connecting me with your PhD and Master's students truly opened a new horizon of research possibilities for my path ahead.

I also want to thank the many other colleagues I have had the pleasure of interacting with: **Shawn**, **Fede**, **Konstantinos**, **Nagarjun**, **Monique**, **Priscila**, **Saman**, **Panchamy**, **Ziyulong**, **Alex**, **Nirvana**, **Sara**, **Dingshan**, **Xiamei**, **Mahsa**, **Elif**, and **Yongqi**. Each interaction, however brief, has enriched my time here. To those who made my life in the Netherlands more joyful and helped me grow in speaking Dutch, **Helen**, **Yash**, **Frank**, **Maria**, and **Kaping**, you may never

read these words, but you made my time here far more fun than it would have been otherwise.

Several people have shaped my future career in Indonesia upon completing the PhD. **Pak Endra** , thank you for offering me the postdoc opportunity. I genuinely appreciate both the trust you have placed in me and for enabling me to do it remotely for several months before I return to Indonesia. **Pak Bayu**, we may have only met once in Groningen, but your advice stayed with me. You encouraged me to find my intrinsic motivation first, to think about where I want to be in the next twenty years, and to let that vision fuel my work with purpose. That guidance gave me the courage to pursue the opportunity I found after completing my PhD. **Pak Yul**, our conversation about the academic landscape in Indonesia when you visited Groningen gave me a new perspective on the opportunities waiting there. **Bang Ridho**, thank you for the career conversations in Groningen. Your wisdom was timely and grounding. **Mas Azka**, what started as a friendship in a lab back in Indonesia has continued here in the Netherlands, and I look forward to working together again back home in Bandung. **Bang Tua**, our conversation at the Singapore conference opened my eyes to what lies ahead in the Indonesian academic landscape and what it truly takes to navigate it well. **Pak Augie** and **Pak Tarto**, thank you for writing recommendation letters for me at the very beginning of this journey, even though things did not go as initially planned. Your support at that early stage is something I will always be grateful for. And to **Ko Sanga**, thank you for the recommendation letter you provided when I needed it most for my scholarship application. I would say that without that letter, I would not be here now.

To my Indonesian friends. First, to my church family at Gereja Kristen Indonesia Nederland (GKIN), thank you for the warmth and community you offered throughout my time here. A special thank you to **Om Stanley** and **Tante Santi** for graciously facilitating our wedding in the Netherlands at the beginning of my PhD. To the married couples of GKIN who have been such wonderful examples for us: **Kak Monique** and **Kak Anton**, **Bang Baput** and **Tante Ita**, **Atin** and **Kristy**, **Teddy** and **Ike**, **Bang Geo** and **Kak Lena**, **Om Erik** and **Tante Yvonne**, **Bang Lucky** and **Tante Lusy**, **Tante Lisa** and **Om Okto**, **Harry** and **Rachel**, **Bang Markus** and **Kak Carol**, **Bang Nelson** and **Kak Pinta**, **Kak Manik** and **Milton**, **Bang Hein** and **Kak Lidia**. To those who helped make our wedding possible: **Nio**, **Kirsty**, **Riando**, and **Claudio**, thank you. Second, to my Indonesian friends whose lives intersected with mine during my Bachelor's at ITB and my time working at Esri Indonesia: **Unggul** and **Ica**, **Gio**, **Rezzy**, **Hanif**, **Suwig** and **Grace**, **Ajeng** and **Tom Bayu**, **Hadi** and **Sarah**, thank you for making me feel as though our friendship continues across countries. Last, to my fellow Indonesian PhD researchers at TU Delft, **Mas Ilham**, **Mas Aga**, **Mas Antra**, **Mas Gilang**, **Fazlur**, **Jerry**, **Yana**, **Panji**, and **Yopi**, it has been a joy knowing fellow Indonesians who were navigating the same journey.

To my family. First, to my parents, **Papa** and **Mama**, both in Bogor and Batam, thank you for always being our support system since the beginning of the PhD and for being ready to visit the Netherlands to take care of **Lucia**. Thank you also to **Tante Rini**, whose help in caring for Lucia gave us the breathing room we needed to finish our PhDs while still being present as parents. The proverb that it takes a village to raise a child has never felt more true than during our time in the Netherlands.

To **Lucia**, thank you for coming into this world while Papa and Mama were doing their PhDs. Some might think having a baby while both parents are doing their PhDs is not a good idea, but you challenged that assumption by being a good and supportive baby. You have also taught Papa to work efficiently so that he can play with you afterwards. Your first and last names

carry the meaning of light, and just as we prayed, you have brought light into Papa and Mama's lives in ways we never expected. Keep shining, my love. May the fear of the Lord guide you in everything you do.

Most of all, to my wife, **Inka**. None of this would exist without you — not the PhD, not this thesis, and frankly, not the version of me standing here today. I still remember telling you that I was content to simply work in industry, that a PhD was not a path I felt I needed to take. But you had a different vision, and you pursued it with such determination that I found myself wanting to come along. What began as accompanying you turned into one of the most meaningful journeys of my life, and I do not regret a single moment of it. We have laughed and struggled together through late nights when the finish line felt impossibly far. But we crossed it together. Thank you for bringing me into this journey, and for supporting me in what comes after this. Without that level of support, I would never have dared to make this decision. Now it is your time to wrap up your thesis. I love you.

Lucas Elbert Suryana
Utrecht, April 2026

About the author

Lucas Elbert Suryana grew up in Bogor (also known as Buitenzorg during the Dutch colonial era), West Java, Indonesia. The city famous of its magnificent Bogor Palace and Botanical Gardens, the oldest botanical gardens in Southeast Asia.

In 2016, Lucas obtained his Bachelor's Degree in Engineering Physics from Bandung Institute of Technology, followed by his Master's Degree in Instrumentation & Control from the same institution in 2019. Following his Master's graduation, Lucas joined Esri Indonesia as a geospatial data scientist. In 2020, he transitioned to academia as a teaching staff member at Calvin Institute of Technology, Jakarta, in the Computer Science Department.

In April 2022, Lucas joined the Department of Transport and Planning at Delft University of Technology as a PhD candidate, fully funded by the Indonesia Endowment Fund for Education (LPDP). His research focuses on the technical operationalisation of the Meaningful Human Control (MHC) concept by developing a methodology to design and evaluate automated vehicle systems that are responsive to human reasons. Building on the insights gained during his PhD, Lucas secured a postdoc position at Bandung Institute of Technology prior to completing his doctorate, where he will continue exploring vision-language model (VLM) reasoning in the context of automated vehicles. This work will be conducted in close collaboration with researchers from the Intelligent Vehicles Group at Delft University of Technology.

Publications

Journal papers

1. **Suryana, L. E**., Calvert, S., Zgonnikov, A., and van Arem, B. Reasons and Principles for Automated Vehicle Decisions in Ethically Ambiguous Everyday Scenarios: The Case of Cyclist Overtaking. *Transportation Research Interdisciplinary Perspectives*, 35, 101787, 2026. *(Included in this dissertation as Chapter 2.)*

2. **Suryana, L. E.**, Nordhoff, S., Calvert, S. C., Zgonnikov, A., and van Arem, B. Meaningful human control of partially automated driving systems: Insights from interviews with Tesla users. *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 113, pp. 213–236, 2025. *(Included in this dissertation as Chapter 6.)*

3. **Suryana, L. E.***, George, A.*, Flipse, L., Calvert, S., van Arem, B., Siebert, L. C., Abbink, D., and Zgonnikov, The Illusion of Control? Linking Behaviour and Perception to Evaluate Meaningful Human Control over Partially Automated Driving. *(In preparation). (Included in this dissertation as Chapter 7.)*

*Joint first authorship. Both authors contributed equally to experimental design, evaluation methodology, analysis, and interpretation of the results.

Conference papers

1. **Suryana, L. E.**, Rahmani, S., Calvert, S. C., Zgonnikov, A., and van Arem, B. A Human Reasons-Based Supervision Framework for Ethical Decision-Making in Automated Vehicles. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 21495-21502). IEEE, 2025. *(Included in this dissertation as Chapters 3.)*

2. **Suryana, L. E.**, Rahmani, S., Calvert, S. C., Zgonnikov, A., and van Arem, B. Evaluating the Alignment of Automated Vehicle Decisions with Human Reasons. In *Proceedings of the 5th International Conference on Robotics, Automation, and Artificial Intelligence (RAAI)* (pp. 734-741). IEEE, 2026. *(Included in this dissertation as Chapter 4.)*

3. **Suryana, L. E.**, Nordhoff, S., Calvert, S. C., Zgonnikov, A., and van Arem, B. A meaningful human control perspective on user perception of partially automated driving systems: A case study of Tesla users. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)* (pp. 409–416). IEEE, 2024. *(Included in this dissertation as Chapter 5.)*

Other publications (not included in this dissertation)

1. Rahmani, S., Neumann, J., **Suryana, L. E.**, Theunisse, C., Calvert, S. C., and van Arem, B. A Bi-Level Real-Time Microsimulation Framework for Modeling Two-Dimensional Vehicular Maneuvers at Intersections. In *Proceedings of the IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)* (pp. 4221-4226). IEEE, 2023.

2. **Suryana, L. E.**, Bierenga, F., van Buuren, S., Kooij, P., Tulleners, E., Scari, F., Calvert, S. C., van Arem, B, and Zgonnikov, A. CARE-Drive: A Method for Evaluating Reason-Responsiveness of Vision–Language Models in Automated Driving. *Transportation Research Part C: Emerging Technologies*. *(Under review).*

TRAIL Thesis Series

The following list contains the most recent dissertations in the TRAIL Thesis Series, For a complete overview of more than 400 titles see the TRAIL website: www.rsTRAIL.nl.

The TRAIL Thesis Series is a series of the Netherlands TRAIL Research School on transport, infrastructure and logistics.

Suryana, L.E., *Operationalising Meaningful Human Control for Reason-Responsive Decision-Making in Automated Vehicles*, T2026/13, June 2026, TRAIL Thesis Series, the Netherlands

Yan, H., *Cycling Speed: Variation and Stability within Rides*, T2026/12, May 2026, TRAIL Thesis Series, the Netherlands

Abhishek, D., *Safe Navigation of Autonomous Vessels in Inland Waterways under Uncertain and Abnormal Operational Condition*, T2026/11, May 2026, TRAIL Thesis Series, the Netherland

Garrido-Valenzuela, F., Pixels, People, Places: *Computer Vision and Image Embeddings for Perception-Aware Urban Analytics*, T2026/10, April 2026, TRAIL Thesis Series, the Netherlands

Spierenburg, L., *Advances in the Analysis of Residential Segregation and Urban Riots*, T2026/, April 2026, TRAIL Thesis Series, the Netherlands

Boot, M., *Evaluating Experiences with Smart Cycling Technologies: Sensor-based evaluations of outdoor cycling experiences with Smart Cycling Technologies*, T2026/8, March 2026, TRAIL Thesis Series, the Netherlands

Wen, X., *Data-Driven Spatial-Temporal Modeling for Bicycle Traffic Prediction*, T2026/7, March 2026, TRAIL Thesis Series, the Netherlands

Wang, Z., *Optimising Performance of Automatic Train Operation on Railway Networks*, T2026/6, March 2026, TRAIL Thesis Series, the Netherlands

Hadi, A.H., *DEM Modelling of Multi-Component Segregation in the Blast Furnace Charging System*, T2026/5, February 2026, TRAIL Thesis Series, the Netherlands

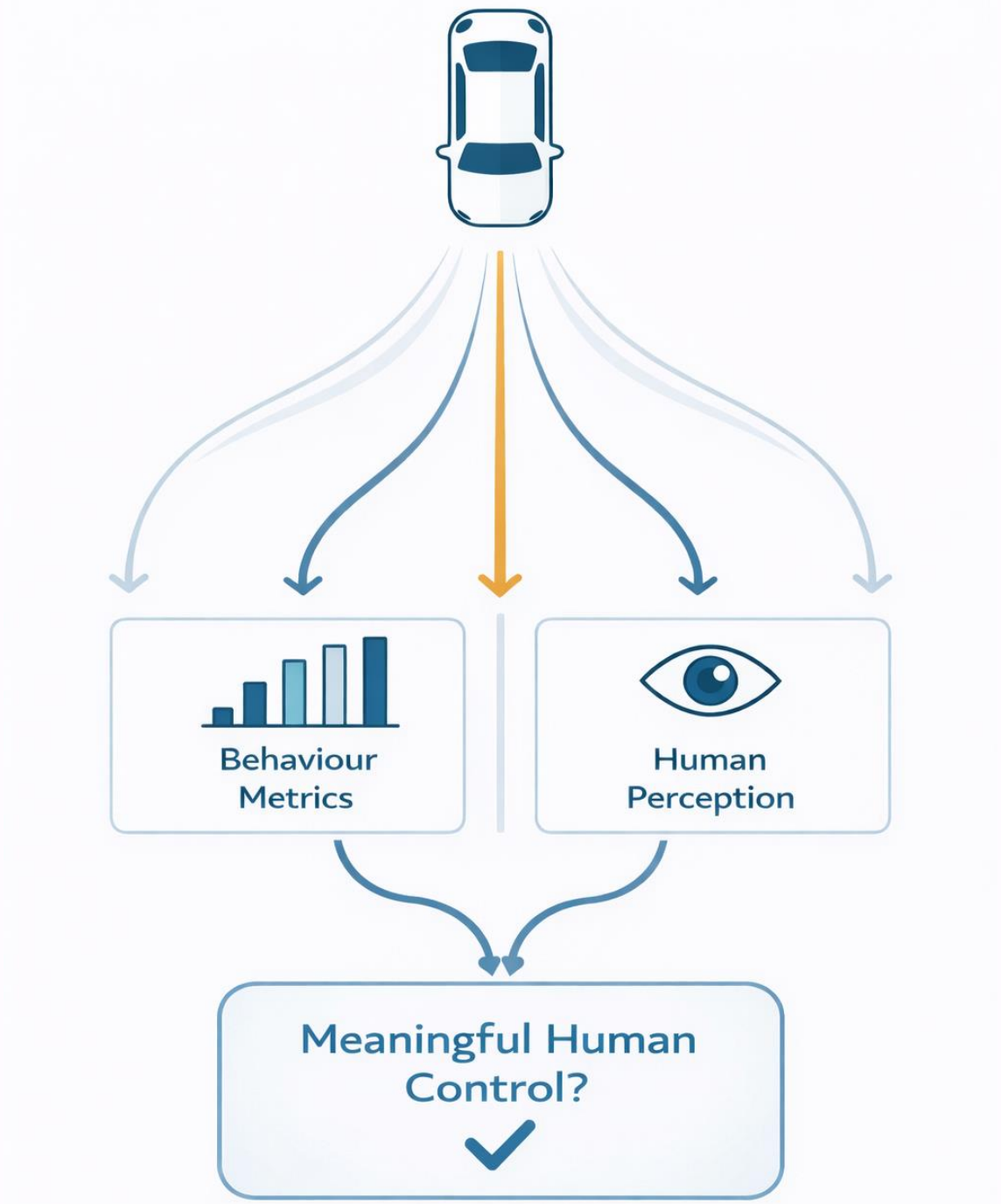
Farhani, M., *Demand Management Strategies for Operations of Shared Mobility Services*, T2026/4,

February 2026, TRAIL Thesis Series, the Netherlands

Yao, X., *Driving Heterogeneity in Traffic Flow Theory: An action-based framework for identification, modelling, and simulation*, T2026/3, January 2026, TRAIL Thesis Series, the Netherlands

Versluis, N.D., *Optimising Railway Traffic Management under Radio-Based Distance-to-Go Signalling*, T2026/2, January 2026, TRAIL Thesis Series, the Netherlands

Jiao, Y., *Proactive Collision Risk Quantification in Multi-directional Traffic Interactions*, T2026/1, January 2026, TRAIL Thesis Series, the Netherlands


Behaviour Metrics
Human Perception
Meaningful Human Control?