A Systematic Implementation and Validation of Distractions in TCI Traffic Models

F. D. Dutruel

TUDelft

A Systematic Implementation and Validation of Distractions in TCI Traffic Models

by

F. D. Dutruel

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday July 15, 2025 at 11:00 AM.

Student number:	5011507	
Project duration:	December 10, 2024 – July 01, 2025	
Thesis committee:	Dr. Ir. S.C. Calvert	TU Delft, Civil Engineering, chair
	Dr. Ir. I. Martínez,	TU Delft, Civil Engineering, supervisor
	Dr. A. Zgonnikov,	TU Delft, Mechanical Engineering, supervisor

TU Delft

Abstract

Human factor models has been an increasingly more popular topic in traffic models. The objective of these models vary, from simulating cooperative driving to understanding the behaviour of distracted drivers. Regardless of these diverse objectives, the reasons motivating these researches boil down to one single reason, safety. By better understanding human behaviour it should be possible to increase the safety of drivers on the road. One model which offers a systematic approach of studying human factors is the task-capability interface (TCI) model, it models the underlying human thought process and uses it as a proxy for other human factors. This has made the model quite successful in replicating various human factors, including distractions. Multiple papers have studied distractions using the TCI model as a tool but they all had their own specific approach to distractions. This leads to the identified gap in literature: how can distraction be systematically modelled in a TCI traffic model.

To fill this gap a distraction framework has been developed. This framework relies on the low-level characteristics of distractions and separates their lifecycle into three stages. These stages are the distraction trigger, intensity and effect. In order to verify if this framework is capable of systematically and accurately modelling distractions it was subjected to a validation test. To this end the distraction framework was incorporated into the Multi-scale model, which was found to be the most fitting TCI model, this resulted in the Distraction model. The new Distraction model was subsequently calibrated with a genetic algorithm to two different datasets with vastly different distraction, a continuous mental-visual distraction and a spontaneous auditory distraction. The results were compared to the results of specialized Distraction models.

The validation test results show that the Distraction model has shown limited improvements over the specialized baseline models and that most of the time its performance is equivalent. To be more specific the Distraction model is significantly better at estimating the headway of drivers compared to the baseline models when calibrating for single drivers. That said when it's used as a calibrated model it loses this edge and its performance is fully equivalent to the baseline models. With these results it can be concluded that the distraction framework functions as intended. Despite the limited amount of different distractions in the validation test it has shown that it is capable of systematically modelling distraction on a similar level as other specialized models. This also shows that the main benefits of the framework are its systematic approach and flexibility and not its performance capabilities.

Preface

This thesis marks the end of my master study at the Delft University of Technology. It has been a challenging journey for me, requiring multiple changes in project direction due to lack of public data and infeasible validations. To this end I would like to thank Hans van Lint who granted me access to the UNINA dataset from i4Driving before the i4Driving project reached its completion and publication. Similarly, I am grateful to Zuduo Zheng for providing me with the CARRS-Q dataset despite having moved on from his work at the Queensland University of Technology. Lastly the guidance on technical and practical matters regarding the datasets offered by Xiaolin He allowed for a quick and smooth integration of the data in my validation tests.

I would like to extend my sincere thanks to the members of my thesis committee at the university for joining me on this journey. In particular to my supervisors Irene Martínez Josemaría and Arkady Zgonnikov, whose insights and contacts allowed me to take the thesis in a direction I didn't know was possible.

I am also grateful to my family, who was always present to offer me their opinions and advice, supporting me morally throughout the completion of this work. Whether this was to help rubber-ducking the logic or proofreading the thesis, they always kept a positive attitude which let me stay motivated to finish this work.

F. D. Dutruel
Delft, June 2025

Contents

1 Introduction

Microscopic traffic simulation is a broad and diverse topic with plenty of different model concepts, scopes and distinct implementations. One specific type of model are models which incorporate human factors (HF). These are a more recent development compared to the older more established traffic models which only answer the questions of how vehicles move from A to B in accordance with all traffic rules. Traffic models which take human factors into account also try to answer one additional question, namely: why and when do vehicles fail to follow the traffic rules (Saifuzzaman & Zheng, 2014).
Incorporating HF in models is becoming an increasingly more popular research topic in the last decade, this is the result of the increasingly automated and digitalized commercial traffic ecosystem (Eltoweissy et al., 2010). The increasing interest in autonomous vehicles and interconnected infrastructure can be seen in nearly every mode of traffic, from bicycles to trucks. These two types of development heavily depend on precise control systems and accurate model predictions. As long as traffic isn't fully automated and contains some level of human intervention, then the models used to predict traffic need to take those human actions and distractions into account. Hence the need for more specific HF models.

With the boom in new HF development there are a number of different theories available defining the implementation of human factors in traffic models (Saifuzzaman & Zheng, 2014). For example, some models aim to alter the perception either by introducing a threshold which needs to be passed in order to detect a change or by changing the perceived stimuli from non-intuitive absolute values to intuitive relative values (Andersen & Sauer, 2007; Wiedemann, 1974). Other types of models introduce subjective risk-taking, these models relax some of the safety limits set on parameters to take certain actions based on the subjective risk perception of the driver (Hamdar et al., 2008, 2015). Each of these theories have their pros and cons and are supported by a limited amount of papers. This thesis will focus on the task-capability interface traffic (TCI) model, a control model which aims to provide a generalized explanation for the human decision-making process. It does this by modelling the human thought process. As a generalized model, the TCI model provides no exact formulations to be used in the implementation of the model, it only provides general guidelines (Fuller, 2011). Furthermore since the thought process is such a low-level human factor it can be used as a proxy to emulate other human factors. This makes TCI models one of the more flexible models when it comes to human factors. With these guidelines multiple implementations have been made covering various facets of HF but one of the aspects of HF which hasn't been fully studied with TCI is distractions. Distractions are known to be a major source of collisions and other critical incidents (Gordon, 2008; Regan & Hallett, 2011). The hope is that by studying distractions on a microscopic scale it will be possible to find root causes for the collisions and to develop mitigations for these causes. Several papers have studied the implementation and effects of distractions in a TCI model and have found promising results (Li et al., 2020; Saifuzzaman et al., 2015, 2017; Van Lint & Calvert, 2018). That said one of the commonalities of these papers is that they only study one specific type of distraction per model with their own specific approach instead of trying to study distractions as a greater whole. This difference in methodology and implementation makes it quite difficult to accurately compare the results of the different papers. This in turn hampers the comparison of different distraction types on driver behaviour. This leads to the topic of this research, namely: The systematic implementation of multiple diverse driver distractions in a microscopic traffic simulations.

This document has the following structure. Chapter 1 is the introduction, presenting the research topic and its relevance. The next chapter is about the research question, sub-questions and the reasoning behind them. Chapter 3 provides a basic literature review used as an introduction to the topic and as a base for the methodology. Chapters 4, 5 and 6 together form the methodology of this thesis. More specifically chapter 4 uses the findings about distractions in the literature review to determine the base framework of the Distraction model. In chapter 5 the found TCI models from the literature are analysed and compared with a multi-criteria analysis to find the best base model for the Distraction model. Chapter 6 covers the methodology for the validation of the Distraction model, including test data, Distraction model implementation and calibration methods. Chapter 7 provides an analysis of the results from the validation tests. Chapter 8 is the discussion about potential weaknesses in the thesis and alternatives which could improve the results some of which fall outside the scope of the thesis. Lastly chapter 9 used the found results from the previous chapters to answer the main research question and provide the final conclusion of the thesis.
In addition to the main chapters this thesis also contains multiple appendixes to provide supplementary material.

2

Research Question

Based on the discussions of existing research and further study into the topic of the task-capability interface model it is clear that there has been limited research into systematically modelling different types of distractions in TCI models. The paper by Saifuzzaman et al. (2015) mentions this gap in their discussion and state that while it is theoretically possible to implement different types of distractions into their model it would require some modifications and recalibration for each distraction making it far from an ideal solution. Knowing this, it's safe to say that distractions in TCI are a gap in the current scientific research.
In order to help, at least, partially fill this gap the topic needs to be translated into a research objective. To this aim a research question is formulated. The main research question is as follows:
"How can a framework for systematically defining different types of distractions be implemented and validated at a microscopic scale in a task-capability interface car-following model?".

Based on the above main question the following sub-questions have been derived:

1. What should the distraction framework properties be in order to support different types of distractions?
2. Which TCI model implementation should be used as a base model to implement the distraction framework?
3. How should the logic and parameters of the base TCI model implementation be modified in order to integrate the distraction framework?
4. Which distraction types and TCI model implementation should be used to validate the proof of concept for the distraction framework?

While the social and scientific relevance of the main question has been made clear in the Introduction and the first section of this chapter, the relevance of the sub-questions hasn't been addressed yet. The following section will briefly address the relevance and reasoning behind each sub-question.

The first sub-question addresses the nature of distractions. In order to implement distractions in simulations a definition of distractions is needed. What are the different distraction types and what are their differences? Knowing this is the foundation to creating a distraction framework, whose purpose it is to quantify distractions. This framework is the main tool which will help answer the main research question. Because of this, the structure and characteristics of the framework will affect the choices made in the following sub-questions.
The second sub-question aims to find an TCI model implementation which is compatible with the framework made in the first sub-question. This selection is then further refined by the modellers' personal preferences and or external circumstances. The selection for the base model needs to be thoroughly researched and well-reasoned since it will have a large impact on the final model's performance. Furthermore if the implementation isn't fully compatible with the framework it will be impossible to validate all the components of the framework. This in turn means that the main research question can't be

answered.
With the distraction framework complete and the TCI model implementation known the two need to be merged. This is the main focus of the third research question, how do you merge two different models without compromising the performance and objectives of either of them? It is important to understand how the model logic and parameter affect the model performance and which parts can be modified without negative consequences. The TCI conceptual model provides various approaches and parameters to influence the decision-making process, these options can be used to model the various effects of distraction. Some are earlier in the model like the task demand and task capability and some occur later like the situational awareness. While influencing these parameters can have similar effects on the final driving behaviour, their details and meanings vary depending on the implementation. By showing the logic used to make these kinds of decisions during the integration step it should be possible to justify the modifications which were made in order to build the final Distraction model. Additionally using this step as a guide it should also be possible to integrate the distraction framework into different kinds of car-following models as long as they meet the requirements set in sub-question 2.
Lastly, the distraction part of the final model needs to be validated. This requires comparing the results to another valid data source and determining to what extent the two are comparable. This is nearly impossible to do since it is a general model which should cover all possible distractions. So in order to completely validate it, all distraction types need to be tested and validated individually. The main issue with this is the limited data availability regarding the different distraction types, if there is no compatible dataset for a distraction type then the model can't be validated for that distraction type. Instead the scope of this thesis will need to be restricted. Since the model can't be validated in its entirety, it will at least need to show a validated proof of concept. The idea is as follows: In order to prove that the general Distraction model is better at representing various types of distractions than specialized Distraction models, the two models will each run two scenarios covering different types of distraction, one of which being the distraction type of the specialized model. If on average the results of the general model are better than the results of the specialized model then it shows that the general model and in turn the distraction framework are a working proof of concept. So in order to carry out this proof of concept a couple of different components are needed, first two different distraction types and dataset and secondly a specialized Distraction model for one of the two distraction types. Finding these two components is the goal of research question four.

3

Literature Review

The task-capability interface (TCI) model is the base model which is going to be expanded in this thesis. In order to do that a firm understanding of the theories used by the model is needed. The first section of the literature review will cover the fundamentals of the TCI model, the various TCI model implementations which will be compared to one another and some of the latest developments in regards to this model.
The second part of the review contains information regarding distractions. It defines what distractions are, gives an overview of the characteristics and some of the interference caused by distractions.

3.1. Task-Capability Interface Model and Risk Allostasis Theory

The task-capability interface model is a model proposed by Fuller (2011) in order to understand the decision-making process of drivers, a complete overview of the model can be seen in Figure 3.1. The core of the model is the task capability comparison. The model assumes that without any active control the vehicle is guaranteed to crash and it's up to the driver to prevent this from happening. These interventions from the driver are called tasks and they happen continuously from origin to destination. Each task has an associated load called the task demand and each driver-vehicle combination has a maximum capability. If the total task demand exceeds the capability of the driver, it means that the driver loses control which results in one of two outcomes: a lucky escape or a crash.

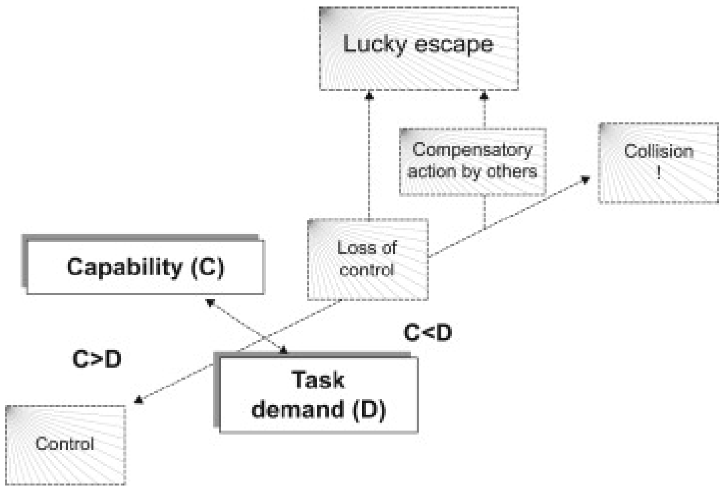

Figure 3.1: The conceptual model for the task-capability Interface model (Fuller, 2011)

This gap between the task demand and the capability is referred to as the task difficulty. The smaller the margin between the two the higher the difficulty. The process of keeping the demand below the capability happens internally and subconsciously hence it's called the task difficulty homeostasis. The

result is that the driver selects actions in such a way as to keep the task demand below his own capability.
A previous work by Fuller et al. (2007) also explores the correlation between task difficulty and perceived risk of actions of the driver. It concludes that there is a strong correlation between the perceived risk of a task and the experienced task difficulty. This link is the basis for the risk allostasis theory, which states that perceived risk can be translated into task demand.
Based on these two theories the control process of drivers in the TCI model is as follows: A driver subconsciously perceives risks associated with the different actions he can take. These risks are translated into demand and compared against the driver's capability. If the demand exceeds the capability then the driver will seek out a different set of actions which doesn't disturb the task difficulty homeostasis. If no such action can be found the driver ends up losing control which can lead to a crash.

3.2. TCI Model Implementation

At its core the task-capability interface model is a theoretical traffic control model. Fuller provides a conceptual model and the reasoning for why it should work but there is no actual implementation (Fuller, 2011). In order to use the model in a simulation an actual implementation needs to be developed for this purpose. There exist multiple implementations of the TCI model each of which have their own pros and cons. The implementations reviewed in this literature study are the following: the modular simulation framework by Calvert et al. (2020) and Van Lint and Calvert (2018), the task difficulty model by Saifuzzaman et al. (2015, 2017) and the fuzzy task difficulty LCM by Li et al. (2020).

3.2.1. Multi-scale Framework

The implementation by Calvert et al. (2020) and Van Lint and Calvert (2018) is one of the more recent implementations which is still undergoing active development. The main aim of this implementation is to have a modular system which uses simple but accurate calculations so that it can be used in larger simulation programs. The framework is split into two major parts, the cognition and control layers. The control layer is made of idealized (collision-free) models, these don't include any human factors in them, an example of this is using IDM+ for the car-following model. The cognition layer is the part of the framework which includes the human factors. The details of the layer's implementation of TCI will be explained below. It's important to understand that the only communication between the two layers is the reaction time, the (biased) perceived stimuli and the driver's preferences. As mentioned before, the framework is made to be modular and each system or subsystem can be replaced with another as long as it accepts and provides the correct inputs and outputs.

The default cognition layer used in the framework is based on the TCI model. Like the TCI model, each driver needs to actively correct their vehicle's course to avoid crashing. A complete overview of the cognition layer can be seen in Figure 3.2 as is explained in this paragraph. The cognition layer is modelled as follows, a driver has a set of active tasks each with their own task demand (TD_i) depending on the current environment. The task capacity (TC) is fixed and isn't expected to change. By dividing the sum of task demands by the task capacity you get the task saturation (TS) which represents how much of a mental load a driver is experiencing. The task saturation can go above 1 but this usually leads to detrimental effects in perception. This task saturation is then converted into situational awareness (SA) according to a quadratic curve, meaning high SA when under moderate TS and low SA when under high or low TS. The situational awareness is then used as a coefficient to define the driver's perception precision. The formulas for each of these steps can be found in appendix C.3 Multi-Scale.

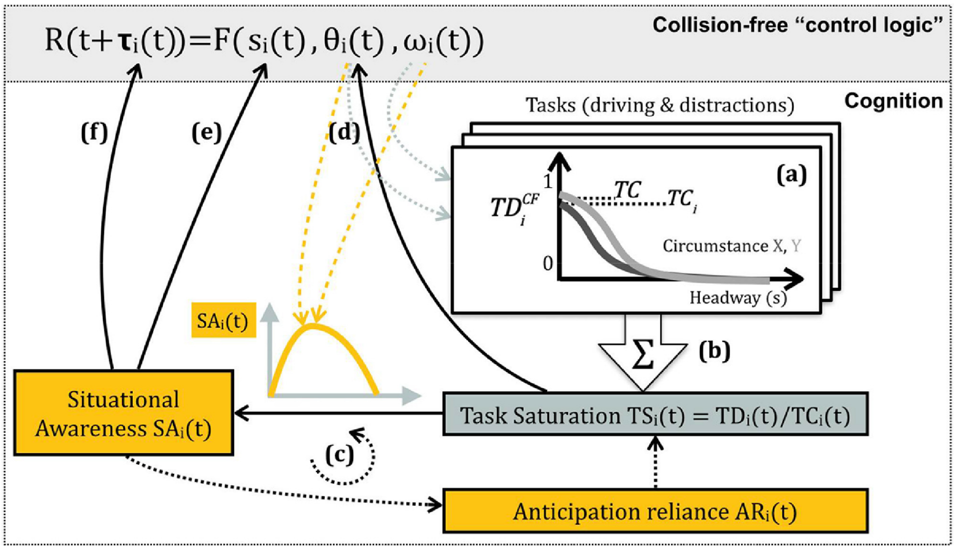

Figure 3.2: Human Factor model simulation framework (Calvert et al., 2020)

Another element seen in the figure above is the anticipation reliance (AR). Anticipation reliance is based on a theory explaining how humans anticipate and how much confidence they have in that prediction. The human thought process can't multi-task, instead during situations like driving where there are multiple tasks that need to be managed in parallel a human places its focus on one task and frequently rotates which task is the centre of attention. This task is called the primary task, while all other tasks at that moment in time are the secondary tasks, which are managed by subconscious anticipation. Hence the concept of anticipation reliance. Based on a driver's experience and the inherent difficulty of a task, the task demand of secondary tasks is reduced by their anticipation reliance, this in turn reduces the total task saturation. Anticipation reliance often has little to no effect on the primary tasks, which usually means that the primary task is the largest contributor to the task saturation.

The Multi-scale framework model is still in active development. It's currently being used as part of the i4Driving project[1] which is a European project aiming to establish a more credible and realistic human road safety baseline. One such development is by Wouter Schakel from TU Delft. His current research is about improving the task prioritization algorithm. The idea is as follows, the physical area around the vehicle is divided into zones. The driver can only observe one zone at a time, as a result, the tasks in that zone are the primary tasks and the remaining tasks are secondary tasks. The way the primary zone is selected is with a Markov chain[2], this should give a realistic model of how humans divide their attention. At the time of writing the thesis this work hasn't been published and will not be used in the thesis.

3.2.2. Task Difficulty Car-Following

As the name suggests, the Task Difficulty Car-Following (TDCF) model by Saifuzzaman et al. (2015, 2017) is a model which aims to implement the TCI concept in car-following models. The idea in TDCF is to include the task difficulty (TD) as a variable in the car-following formulation in order to accurately capture human factor influenced driving and normal driving in a singular model.

The task difficulty as defined by Fuller (2011) is the gap between the task demand and driver capability. While this definition is clear on a theoretical level there is no explicit formulation for capability or task demand. In order to counteract this problem Saifuzzaman uses research by Lewis-Evans et al. (2010) which proves a correlation between desired headway, actual headway and task difficulty. The research states that if the headway is below the desired headway then the driver finds the driving task more difficult, and similar results for the opposite situation. With that knowledge the TD is formulated

[1] https://i4driving.eu/
[2] https://en.wikipedia.org/wiki/Markov_chain

as follows:

$$TD_n(t) = \left(\frac{V_n(t-\tau_n')\tilde{T}_n}{(1-\delta_n)S_n(t-\tau_n')}\right)^{\gamma} \tag{3.1}$$

To those familiar with CF models and their formulas they will recognize three new parameters. The first two, δ_n and γ, are directly related to the task difficulty while the third parameter, τ_n' is a parameter which is used in all the CF formulas and not just the task difficulty.
To be more specific δ_n is the risk perception parameter, this parameter represents how aware a driver is of difficult situations and or distractions. This parameter can be both positive and negative. When positive the driver will be more careful in low headway situations and when negative they will be more aggressive.
The γ is the sensitivity parameter, it exponentially increases or reduces the task difficulty. The paper shows that the parameter doesn't have a big influence on the equilibrium solution of the final CF models. It's mostly used to create a better fit for individual drivers but the discussion and analysis of this parameter is largely ignored in the paper. That said this paper does have a real world counterpart, namely the sensitivity to certain types of distraction, these sensitivities are unique to each driver so in that sense this parameter does behave how it should.
Lastly, there is τ_n' which is the modified reaction time, it is calculated as follows $\tau_n' = \tau_n + \varphi_n$. This is how the model incorporates the effects of distractions. The φ_n is the increase in reaction time due to external situation. This parameter is then applied to the complete CF formulas.
The complete formulation of the model can be found in C.2 TDIDM. Additionally an overview of these parameter and variables can be seen in the flow chart bellow:

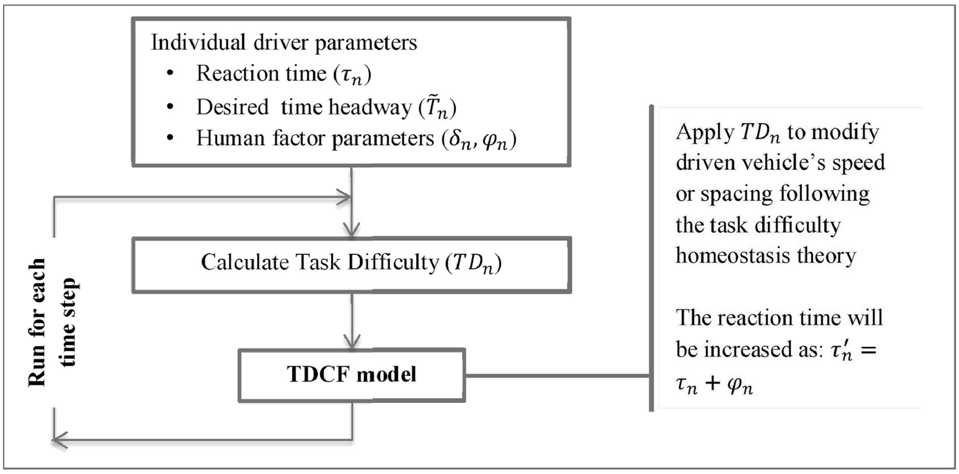

Figure 3.3: Task difficulty car-following conceptual model (Saifuzzaman et al., 2015)

To test the effectiveness of the TDCF theory Saifuzzaman et al. (2015) applied it to two well known and performant models. These are the Gipps model and the IDM model, the new models are called TDGipps and TDIDM. The task difficulty variable only affects either the speed or the spacing and it makes sure that the model remains consistent with the task difficulty homeostasis.
Thorough analysis and validation with simulator data shows that the new models are better than their base counterparts.
This includes having stable equilibrium speed-spacing equations in all possible cases, unlike the base models. The models have significantly better speed and spacing accuracy when comparing it against the simulator data compared to the base models. This is true for both normal driving and distracted driving cases. Lastly the models are also less sensitive to calibration errors. While the base models can simulate a variety of different driving behaviour these aren't necessarily realistic for all drivers, the TD models avoid these mistakes with the self-correcting factor of the task difficulty.

Despite the positive results, there are still a few areas that the author points out which need improvement. Furthermore some of the base assumptions made in the paper conflict with theories in literature. One of the main problems which the author point and which are also relevant to this research is that

the humans parameters δ, γ and φ are fixed at the start of the simulation. While γ is unlikely to change during a trip, δ and especially φ can change during one trip. The author does mention that these issues could be solved if the distractions were simulated in a more dynamic way which would allow appropriate values to be picked from a calibrated matrix.
Another interesting point which the author noticed during calibration is that the model could provide a better fit if the addition reaction time φ can be negative, which would mean that the driver pays more attention to the road while having a distraction present. The author prevented this from happening due to one of the assumptions made at the beginning, namely that distractions can only worsen driving behaviour. This assumption contradicts some of the assumptions made by Van Lint and Calvert (2018) which says that distractions can either improve and worsen driving performance.
Lastly a problem which the author doesn't mention but is probably aware of due to the selected scoping is the fact that the author chose to only cover one type of distraction. We know from literature that there are various types of distractions and that they can have various types of effects on the driver. Despite this the author chose to only cover one distraction, handheld phone calling, and only one effect, increased reaction time. In order to properly include distractions in a model the model needs to be able to cover multiple types of distractions and multiple types of effects, this would likely necessitate multiple δ, γ and φ for each distraction type making the model difficult to calibrate and validate. Hence the reason for a limited scope.

3.2.3. Fuzzy Task Difficulty Longitudinal Control Model

The Fuzzy Task Difficulty Longitudinal Control Model (FTD-LCM) model by Li et al. (2020) is an implementation of the TCI framework which tries to improve upon existing TCI model implementations and merge their benefits. The two main inspirations for the FTD-LCM models are the Multi-scale model (Van Lint & Calvert, 2018) and the TDCF model (Saifuzzaman et al., 2015) which are discussed above.

The authors of the FTD-LCM identify a total of two problems with the previous models and three gaps which are filled with this model.
The first found problem is as follows, the approach of using the deterministic mathematical equations to estimate a driver's experienced task difficulty fails to take the uncertainty and ambiguity of a driver's perception and decision-making into account. To this end the authors propose to use fuzzy logic to change the model from a deterministic model into a stochastic model. This change should help better capture the fuzzy nature of the human brain than the existing models.
The second issue is with the choice of car-following model in the Multi-scale model. The Multi-scale model uses the IDM+ model in its test, while this is a commonly used model it isn't highly motivated why this model is used. Instead of using the IDM+ model the authors motivate that the Longitudinal Control Model (LCM) should be used instead since this model is based on a more sound theoretical background (Ni et al., 2015).
These two problems and potential solutions can be linked to a total of three research gaps. First the requirement for a more desirable underlying CF model for the framework of the TCI model. Second the ignorance of the fuzziness of human brains when modelling driver's task difficulty. Lastly the demand for more inclusion of human factors into the LCM model.

In terms of model structure the FTD-LCM model uses the Multi-scale model as base, it keeps the overall logic and layout of the two framework layers but modifies their internal components. For the control layer the model uses the LCM instead of the IDM+ model, the model makes no other modification to either the LCM model or the control layer. As for the cognition layer it is in this layer that the fuzzy logic is implemented. The overall structure of this layer remains largely the same as seen in Figure 3.2 but the calculations between each element are different. In terms of fuzzy logic the model uses two different functions. First a Gaussian membership functions to convert a numerical value into a fuzzy linguistic label value, fuzzy value for short, and back. Secondly the mapping functions which map one (set of) fuzzy value(s) to another fuzzy value.
With these two functions the FTD-LCM model modifies the cognition layers as follows. First the numerical spacing, ego speed and leader speed are transformed into relative position (RP), relative velocity (RS) and current velocity (CV) fuzzy values. These three fuzzy values are then mapped to a fuzzy task demand (TD) value which is then transformed into a numerical TD value. The task saturation (TS) calculation remains untouched so it remains a mathematical formula but once the TS is calculated it

is again transformed into a fuzzy TS value. The fuzzy TS values is then mapped to a fuzzy situational awareness (SA) value which is again transformed into a numerical SA value to calculate the error from the optimal SA. This SA error is then used to bias the numerical perception of the driver's surroundings and also delay the driver's reaction time.
In total the whole cognition layer has to process six membership functions from start to end which adds a lot of statistical uncertainty.

As for performance of the model does a fairly ordinary calibration-validation test with the NGSIM dataset[3].
The trajectories are all calibrated individually with a genetic algorithm whose objective function is based on the difference in acceleration of the simulated and real vehicle. This calibration is run both for the FTD-LCM model and the LCM model. The result of this calibration shows that for all the used trajectories the LCM model has a lower median value for the objective functions but the FTD-LCM model has fewer bad outliers and more good outliers; the mean objective value of both models is not given.
As for the validation test, the authors used 5 out of 262 trajectories for validation while the rest was used for calibration. The validation test is run with a model that uses the mean values of each calibrated parameter and the error again measured based on the acceleration. Overall the FTD-LCM seems to provide slightly lower error values than the LCM model but no t-test is used to prove that this difference is significant.

This leads to the issues of this model. While the provided theory and modifications to the Multi-scale model are sound and well reasoned the results are questionable. Using the acceleration as parameter to calculate the error could be problematic since the NGSIM dataset is a set of observed trajectories, this means that all the acceleration values are calculated based on the position and not measured hence they are imprecise. Furthermore when manually calculating the t-test for the validation results it results in a p-value of 0.85. This is too low to ensure that the difference between the two error measurements are significant, so it can't be concluded that the FTD-LCM is significantly better. The cause for the failed t-test is likely to be the small data batch size for the validation test. If a t-test with a larger batch is run and still fails to pass then it likely means that the model doesn't provide a significant improvement over the existing LCM model.

3.3. Distractions

The topic of the thesis is driver distractions. Most people would have a reasonable default understanding of what these represent in general but in order to work with distractions scientifically we need a clear definition of distractions. The primary source of information comes from a book chapter written by Regan and Hallett (2011), this information is additionally cross-referenced with newer papers about driver distractions.

When looking at driver distractions it's important to know that driver distractions are only a smaller section in a much more complex topic called driver inattention. The definition used by Regan and Hallett (2011) of driver inattention is: "insufficient or no attention to activities critical for safe driving". At first this might sound a lot like driver distractions but there is an important distinction. In driver distractions the attention of the driver is drawn away specifically by a competing activity.
The other categories of driver inattention are: restricted attention caused by a physical obstruction (for example by the front windscreen pillars), misprioritized attention caused by excluding one critical activity due to spending too long on another critical activity, neglected attention caused by a driver neglecting a critical activity and cursory attention caused by a driver only giving cursory attention to a critical activity.
While misprioritized attention and driver distraction might resemble one other the important distinction is that misprioritized attention is caused by a critical activity while distraction are caused by non-critical activities. That said the non-critical activity can still be driving-related, like planning your route on a strategic or tactical level.
In the definitions above the term critical driving activity has been used. But what exactly qualifies as a critical driving activity? The currently accepted definition of critical driving activity is that it's a sub-

[3]https://data.transportation.gov/stories/s/Next-Generation-Simulation-NGSIM-Open-Data/i5zb-xe34/

set of primary driver tasks. Primary driver tasks are activities which are essential to driving safely (e.g. changing gears, using indicators), note that this definition of primary task is different from the definition of primary tasks in the Multi-scale model. The complementary set of primary tasks are the secondary tasks which are all tasks which aren't a primary task (Young et al., 2019). Critical activities differ slightly from primary tasks in that a critical activity is state and environment-dependent. For example, using indicators to enter a highway is a primary task but on an empty road it might not be considered a critical activity; often times when analysing data sets the critical activities are hand-picked (Cuentas-Hernandez et al., 2023). This distinction between primary and critical is a topic that the scientific community is still debating and hasn't been able to resolve so far. This means that critical activities are ultimately model and implementation dependent.

As stated above, driver distractions are caused by other activities which are not critical to safe driving. These competing activities are quite diverse in nature. One of the most notable set of characteristics is whether the distraction is driving-related or not, internal or external to the vehicle, voluntary or involuntary. For example, a driving-related voluntary internal distraction could be looking at your navigation software to see how far away the next turn is.
Another set of characteristics of distractions is the sensory modality. Critical information needs to be observed before one can react to them. Humans detect through their senses but these senses can also be distracted and fail to properly observe critical information in time (Victor et al., 2008). This means that nearly all distractions fall under one or more of the five major senses. These are subsequently called: visual distraction, auditory distraction, olfactory distraction, gustatory distraction and tactile distraction. One special type of distraction which doesn't fall under these distinctions are internal thoughts. These are common enough that they are defined as a sixth category called internal distraction or cognitive distraction.
The characteristics listed above don't specify how distracting a competing activity can be, instead they help describe if and when a driver will be affected by this distraction. Drivers have a certain level of self-regulation when it comes to getting distracted. Some drivers are more easily influenced by distractions with certain characteristics while others are able to completely ignore some types of distractions. This self-regulation is known to be correlated to a number of driver characteristics, the most prominent being: driver age, gender, experience, state (e.g. drowsy, drunk, upset) and familiarity with the distraction (Young et al., 2008). In conclusion the parameters which decide whether a driver gets distracted are the distraction task demand, the critical driving task demand and the driver's sensitivity to the distractions characteristics.
When a driver succumbs to a distraction it usually results in interference to the critical driver activities. This interference leads to a number of outcomes but these are usually split into two categories. Interference that "manifests" and is observable (e.g., lane excursion) or is "intrinsic" (e.g., loss of situational awareness) and unobservable. Having said that an interference isn't necessarily all negative. Some studies have shown that distractions can also improve driving behaviour. An example of this is shown in an experiment performed by Liang and Lee (2010), one of the results of this experiment is that a cognitive distraction resulted in an improvement in lane keeping while decreasing situational awareness.

Now that the definition and characteristics of distractions have been cleared up you might wonder why driver distractions occur in the first place. While the question of why is closely related to when and where it isn't essential to know the answer for making a model about distractions as long as when and where are answered. Answering the question of why falls under the domain of neuroscience and psychology and won't be discussed in this paper.

While not the focus of this research it is good to know how driver inattention is handled by the Multi-scale model. The Multi-scale model follows the TCI model guidelines which means that it implements human factors by modelling the thought process of the driver. This model doesn't explicitly cover driver inattention or driver distractions. That said, the simulation displays effects resembling driver inattention, mainly misprioritized and cursory attention. This is caused by the task prioritization of the model. Since only one task can be the primary task it automatically results in driver inattention despite never being explicitly specified in the model (Calvert et al., 2020).

4

Distraction Framework Definition

Distractions come in many different forms, some quite common and others rarely seen. There have been multiple studies which try to characterize distractions, each with their own approach. These studies usually come to similar conclusions but none are perfect and each approach leaves some gaps which are difficult to explain. To see a more in-depth review of the different characteristics of distractions see chapter 3.3 Distractions.
This chapter aims to answer the first sub-question of the thesis. Namely to find which distractions types exist, what are their differences and how should they be included in the distraction framework.

4.1. Distraction Characteristic Level

The research done in chapter 3.3 Distractions shows that there are multiple ways of grouping different types of distractions together by their characteristics. Some of these characteristics used to group the distractions are considered more low-level while others are a high-level characteristic. Some examples of low-level characteristic grouping are sensory modality or internal/external distractions. Some examples of high-level grouping are conversation, reading road signs or eating. The first step is determining on which level of characteristic the framework should operate. Each has their advantages and disadvantages.
Low-level groupings allow you to have more complex distractions with multiple characteristics. Each of these characteristics can then affect each component of the model separately. This would allow the model to address complex simulations where different distractions interact with one another without interfering. Additionally if in the future a new high-level distraction is found, which can be built out of the implemented low-level characteristics, then it can be implemented without recalibrating the whole model. The main disadvantage of implementing low-level distraction is that a “real” distraction is always composed of multiple low-level characteristics, which means that multiple low-level characteristics need to be implemented in order to correctly simulate a distraction (Regan & Hallett, 2011). This is quite a bit of work, and it would be difficult to validate each low-level characteristic independently of each other. There has been some theoretical research trying to analyse high-level distraction characteristics and split them into low level characteristics but they usually only cover a few distraction types (Engström et al., 2013). This also leads to the second point, namely nearly all data which measures distractions, measures the distraction with high-level characteristics and not low-level characteristics.
On the other hand if the model implements high-level characteristics it will not be nearly as flexible and detailed as a model with low-level characteristics. The main issue with this is that if a new type of distraction needs to be implemented it will probably not be part of the high-level characteristics which were implemented. The main advantage of high-level characteristics is that it’s easier to classify the different distractions with high-level characteristics instead of low-level characteristics. As a result there is also a lot more data available about high-level characteristics.
For the implementation of this framework the level of distraction characteristic abstraction will be the low-level characteristics. The main reason for this is the flexibility and expandability. Since the goal of this thesis is to make a model which could support all types of distractions there is a need for this flexibility and if high-level characteristics were chosen then it would be difficult to validate the model for

all types of distractions.

4.2. Distraction Framework

With a better understanding of distractions, the core theory of this thesis can be made. The distraction framework is responsible for quantifying distractions and deciding how much of an impact a distraction has on a driver. This framework is inspired from existing literature but it is mainly this thesis' interpretation of distractions.
The distraction framework can be divided into three stages, the distraction trigger, distraction intensity and distraction effect. Each of these stages represents an important part of the distraction's lifecycle. These steps are in chronological order and each step refines the distraction from global distraction parameters to personal driver distraction parameters. An overview of the framework can be see in figure 4.1.

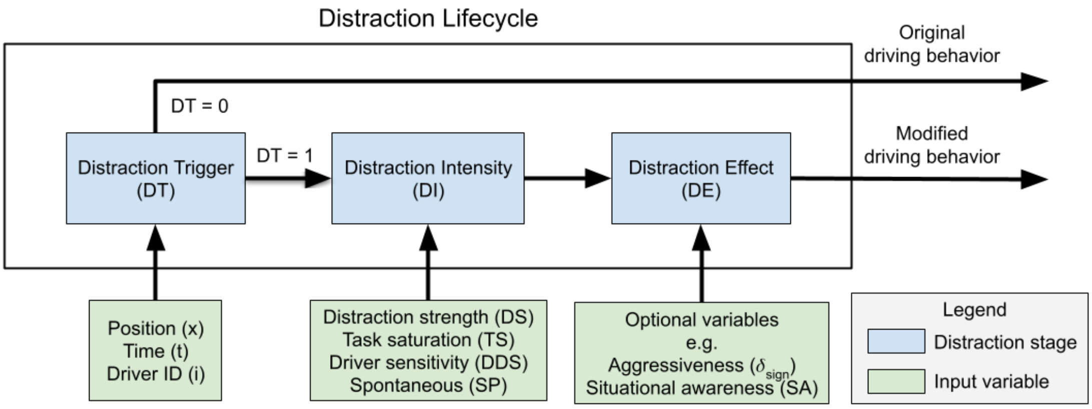

Figure 4.1: Task difficulty car-following conceptual model (Saifuzzaman et al., 2015)

4.2.1. Distraction Trigger

The distraction trigger is the first step in the framework. Its purpose is to define when, where and to whom a potential distraction applies. This step aims to provide an easy way of encoding this information and alert the relevant vehicles in the model. Which kinds of effects can occur or how strong the distraction is aren't taken into account at this stage, as this will be covered in the later stages.

In order for the model to know whether a distraction should apply to a vehicle at a certain time the model can make use of the trigger function. This function is defined as follows:

$$DT_{i,j} = f_{DT,j}(x_i, t, i) \tag{4.1}$$

in which $DT_{i,j}$ is a boolean value which denotes whether distraction j applies to vehicle i; f_j is the specific trigger function of distraction j; x_i is the position of vehicle i, either in absolute coordinates, lane link coordinates, or other coordinates depending on the specification of the simulation. Lastly the subscript and variable i represents the vehicle and t represents the current time of the simulation.
The exact details and formulation of the trigger function depends heavily on the type of distraction modelled. Not all distraction types make use of the provided variables, some would only need a simple condition check on one of the variables while others could require a more complex combination of the variables.

In order to provide more clarity regarding the trigger functions some examples are given below:
A distracting billboard: $f_{DT,j}(x_i, t, i) \rightarrow x_i < x_{billboard}$
A phone call: $f_{DT,j}(x_i, t, i) \rightarrow t_{start} <= t < t_{end} \wedge i = i_{callee}$
An emergency siren: $f_{DT,j}(x_i, t, i) \rightarrow abs(x_{siren}(t) - x_i) < X \wedge t_{start} <= t < t_{end}$

4.2.2. Distraction Intensity

Now that we know if a distraction triggers for a driver we can look at how the driver will react to the distraction. From literature we know that there are quite a few factors at play which determine how strongly a driver will react to distractions, see 3.3 Distractions. That said most of these are minor factors and difficult to measure. So for practical purposes the factors, which determine the level of reaction from the driver, can be reduced to the following: the inherent distraction strength, the driver's task saturation, the driver's distraction sensitivity and lastly whether the distraction is spontaneous or not. Literature also shows that depending on these factors there can be a wide range of responses. Some drivers might completely ignore a distraction if their sensitivity is low enough or if their task saturation is high enough. Whereas in other cases some spontaneous distractions are momentarily impossible to ignore, at first occurrence, but quickly subside.

The purpose of the intensity step is to transform these four parameters into a unique distraction intensity over time graph for the targeted driver. This can be expressed with the following formula:

$$DI_{i,j}(t) = f_{DI,j}\left(DS_j, TS_i, DDS_i, SP_j(t)\right) \tag{4.2}$$

with $DI_{i,j}(t)$ as distraction intensity, DS_j as inherent distraction strength, TS_i as task saturation, DDS_i as driver distraction sensitivity, $SP_j(t)$ as spontaneous function and lastly subscripts i and j as driver and distraction specificity. The specific distraction intensity is expressed in the same unit and range as the TCI model's task demand. This means that the exact values are dependent on the used TCI implementation, that said implementations typically model the TD between 0 and 1 so this will usually also be the case for the specific distraction intensity. The distraction strength denotes the base strength of the distraction, unlike the specific distraction intensity this value is not in the same range as the task demand. Instead it's a positive number, with larger numbers representing stronger distractions. The task saturation also known as task difficulty in the Fuller TCI model, is the ratio between total task demand and task capacity. The driver distraction sensitivity ranges between 0 and 1 and represents how easily a driver is distracted, with 0 being not distractible and 1 being easily distracted. Lastly the spontaneous function is a binary value which denotes if the distraction is spontaneous or not.

The exact formulation of the distraction intensity function depends on the chosen TCI implementation and the type of distraction but in general the following correlations are expected. A higher distraction strength correlates to a higher average specific distraction intensity with no effect on the duration. In contrast a higher task saturation correlates to a lower average specific distraction intensity and or a shorter distraction time. The driver distraction sensitivity acts in the opposite manner to the task saturation. The higher the sensitivity the stronger the average specific intensity and longer the distraction. And lastly if a distraction is spontaneous it usually leads to a peak in distraction intensity at the start and no other changes.

The three examples from the distraction trigger are continued for the distraction intensity. These examples assume the use of the Multi-scale model as underlying model.
A distracting billboard: $f_{DI,j}\left(DS_j, TS_i, DDS_i, SP_j(t)\right) \rightarrow DS_j * DDS_i * min(1,\ 1 - (TS_i(t) - TS_{crit}))$
A phone call: $f_{DI,j}\left(DS_j, TS_i, DDS_i, SP_j(t)\right) \rightarrow DS_j * DDS_i * (1 + SP_j(t))$
An emergency siren: $f_{DI,j}\left(DS_j, TS_i, DDS_i, SP_j(t)T\right) \rightarrow DS_j * DDS_i * max(1,\ 1 + (TS_i(t) - TS_{crit})) * min(0.7,\ \frac{abs(x_{siren}(t) - x_i)}{X})$

4.2.3. Distraction Effect

With the driver reaction over time known it only needs to be transformed into an effect on the driving performance. This is the objective of the last step, distraction effect. The literature review shows that there are multiple changes possible in driver performance due to distraction. These are usually negative changes but in some counter-intuitive cases it can lead to a positive effect on driving performance. This shows that the framework needs to be capable of capturing both positive and negative effects on driving performance. Since this framework doesn't provide a car-following model but is instead expected to be integrated into a CF model it can't make direct changes to the driving logic. Instead it can alter the inputs of the CF model or slightly alter some minor parts of the logic. In order to accomplish this, four common input variable categories were identified from TCI CF models and from the conceptual TCI model, see 3.1 Task-Capability Interface Model and Risk Allostasis Theory and 3.2 TCI Model Implementation. These are the reaction time, task demand, driver personal preference and situational awareness. A distraction effect doesn't have to influence all of these variables, instead it might influence one or two of them depending on the nature of the distraction.

The first two variables need little explanation. Distracted drivers tend to react slower to their surroundings. Furthermore a distraction is a perfect example of a secondary task, and all tasks have some sort of task demand. Depending on the chosen model implementation these two might be everything that is needed to achieve the desired effects of a distraction. That said some distractions might require some more fine-tuning. This is where the last two variables come into play. By modifying the driver personal preferences and situational awareness the distraction can have a more direct control over the driving performance. In TCI models these two variables are usually controlled by the task saturation but in some circumstances a distraction might have an amplified or opposite effect on these two variables than the task demand would usually provide.

5

TCI Multi-criteria Analysis

The distraction framework setup in the previous chapter is a conceptual model, due to this it can't be easily validated on its own. Instead it needs to be integrated into a TCI car-following model which can then be validated with the use of car-following data. To get an appropriate TCI model there are two approaches, either make a TCI model from scratch, this new model could take inspiration from existing TCI model implementations, or use an existing TCI model implementation. Both of these approaches are equally viable, and both approaches need a good understanding of existing TCI models.
This chapter is dedicated to satisfying this need. The goal of this chapter is to review and compare various TCI models to help make a final choice regarding which model approach will be used. This is done with the help of an informal multi criteria analysis (MCA). First a list of criteria is provided, and second the models are compared, and their pros and cons are pointed out.
There are a total of three models, which will be evaluated in this section. These are the Multi-scale Framework, the Task Difficulty Car-Following model (TDCF) and the Fuzzy Task Difficulty Longitudinal Control Model (FTD-LCM). One particularity of note is that two of the three models don't have a formal name. The names used in this thesis are made for ease of reference and are based on the titles of their respective papers.

5.1. MCA Criteria

In order to evaluate and compare the various models systematically a set of criteria is established. These criteria contain both formal model function related criteria and personal preference criteria. The following are the used criteria. Their reason for being included and their grading method are explained below.

1. Model complexity
2. Model flexibility
3. Computational complexity
4. General human factors
5. Existing distractions
6. Author contact
7. Open source

The first criterion is the model complexity. More complex models are typically harder to work with but can give more nuanced results. Furthermore a complex model is also harder to validate and calibrate, resulting in more time spend on those steps.
The goal is to have a model which is complex enough that it can have enough variation in drivers and results while not so complex as to be difficult to interpret and understand. A method of evaluating a model's complexity is by looking up a conceptual model or flow chart of the model. The more steps,

inputs and interconnections there are the more complex the model is.

The second criterion is model flexibility. This refers to how easy it is to modify a model while still retaining its (partial) original function. This brings the following question: “Can the model be modified without requiring a complete revalidation and recalibration.”. Some models are designed to be more robust or modular. They are made in such a way that some parts can be modified or replaced without it having negative consequences on the rest of the model. In our case this is considered a good property since it means that the model will require less work to be validated once the distraction framework has been implemented.

In third is the model’s computational complexity. This has to do with the model’s computational time. The higher the computational complexity the slower the model will be to complete a simulation scenario. This in turn means slower or fewer tests during the model design process and during the verification, validation and calibration phase.
The model’s computational complexity can be evaluated in one of two ways. Either run a standardized test scenario on the model and see how long it takes to complete the scenario. This requires a fully coded model implementation and a pre-validated test scenario which can be time-consuming to build if it isn’t already available in the early stages of model design. Another method is to evaluate the number of complex maths calculations which need to be made and the number of feedback loops taken in a single time step. The fewer feedback loops and maths operations the better the computational complexity.

The fourth criterion is the general human factor inclusion. As explained before the goal of these more advanced models is to more accurately model human drivers, this is done with human factors. If the model already takes into account multiple human factors which each have their own effect on the model it might be hard to evaluate the effect of distractions in the model. On the other hand, a model which has no possibility to include human factors will result in a very robotic way of driving and distractions might have a much larger impact on the results than they would have in reality. This criterion has an overlap with the first criterion but evaluates a specific aspect of the model more in-depth. A positive score is given if the model can implement multiple human factors. Ideally these human factors are modular and can be enabled or disabled at will for the various drivers and tests but if that is not possible it is typically considered better to have human factors than not.

In fifth is the existing distractions, this refers to which types of distractions are already or can be implanted into the model with little to no effort. This is again a more refined version of the fourth criterion. This criterion covers two things, first does the model already implement distractions and secondly how many different types of distractions could this model support. If the model already has a method or framework to implement distraction that could be quite beneficial to the thesis. Having a standardized way in the model to implement distraction would help with testing and with future development of the model since everyone knows what to expect from the already implemented distractions. The second part of the criterion looks at the available variables and data inside the model to determine which types of distractions can be implemented. Most distractions require some specific elements to be present in a model to be correctly implemented, for example, distractions by passengers can only be implemented if the model knows if there are passengers in the vehicle.
The score for the criterion is positive if the model already implements various distractions and if it’s capable of implementing additional distractions.

The sixth and seventh criterion are more subjective than the first five. These criteria don’t evaluate the functionality of the model, instead they represent the personal preference of the author. In sixth is the model author contact. If the original author is easy to reach and is willing to respond to questions regarding the model it makes model development a bit easier since a third-party view can be requested when development is stuck. In Seventh, is whether the code and data are open source. If the code is open source, then the model code does not need to be completely rewritten which saves a lot of time and provides some easy testing opportunities during model development. Additionally, if the data are also open source, then there already exists a complete data and result set that can be used in testing when implementing model changes.

5.2. Model Comparison

Using these criteria each model can be evaluated and compared. The results of this comparison can be seen in table 5.1. A more detailed evaluation of each model can be found in appendix A MCA Model Evaluation. The model which overall performed the best is the Multi-scale model. Overall this model tends to be the most balanced, providing good results in all areas while excelling in flexibility. These components are exactly what is needed for the distraction framework so the Multi-scale model will be used as base model going forward. A more detailed review of the individual criteria can be found below.

Criteria	Multi-scale	TDCF	FTD-LCM
Model complexity	+	++	0
Model flexibility	++	-	0
Computational complexity	+	++	-
General human factors	+	+	++
Existing distractions	+	0	+
Author contact	++	+	+
Open source	++	+	-

Table 5.1: Multi-criteria analysis results of the three reviewed TCI models.

Overall there are two conclusions which can be drawn from the above results. First the models do indeed accomplish what they set out to do and secondly none of the reviewed models are built to specifically handle distractions.
When looking at the results it is noticeable that each of the models excels in a specific area. The Multi-scale model, which is built to be modular and serve as a base for other models, has excellent flexibility while providing good results in other areas of interests. The task difficulty model, which aims to expand upon the most commonly used and simple models, has an extremely fast computational speed and a lower level of complexity compared to the others. Lastly the Fuzzy task difficulty model, which aims to improve the human factors by adding more human-friendly inputs and fuzziness in calculations, has the best human factors.
That said while all models have a method of supporting secondary driving task and distractions none of the models focus on that aspect of driving. None of the models give guidelines on how these should be modelled or when these should occur.

6
Validation Methodology

This chapter covers the various methods and tests used in this thesis in order to help prove the usefulness of the new model when it comes to distractions. As mentioned in chapter 2 Research Question, it is difficult to validate the Distraction model for all types of distractions, instead the tests are made to show a lack of performance in previous models under specific conditions which the new model improves upon. To this end this chapter is divided into three parts. Firstly, how is the validation test designed. This part delves into the data which are used, the TCI model which serves as baseline and the assessment criteria. The second part of the chapter goes into the details of the Distraction model implementation. The Distraction model has mostly been theory until now and will be fully implemented in this section. Lastly, the method of model calibration is covered. Which method is used and what are its benefits over other methods.
Each of these sections contains additional minor tests and additional statistics which are directly presented in this chapter. That said the results of the main validation experiment and analysis of the results are presented in the next chapter.

6.1. Validation Scenario

In order to validate the model a detailed validation experiment needs to be conducted. This section will explain the details of this validation experiment. First the proof of concept is explained. After that the chosen distractions and datasets are explained and reviewed. In third the chosen baseline TCI models are presented. In fourth the assessment criteria are explained. Lastly the two validation tests are explained.

6.1.1. Proof of Concept for Validation

As explained before, the distraction framework made in this thesis won't be fully validated for all types of distractions in this thesis. Instead a partial validation is made which serves to show that the distraction framework is capable of predicting different types of distractions and should be capable of implementing all types of distractions.

To this end a couple of different types of datasets and models are needed. First there is a need for two different types of TCI models which are each capable of modelling a different type of distraction. These two models will be referred to as model A and model B in this subsection. Furthermore two datasets which contain distractions matching the models are needed, these will be called data A and data B.
The idea for the partial validation or proof of concept validation is as follows. If on average the Distraction model is better at modelling the two different types of datasets than the distraction specific models then it proves that the distraction framework is capable of implementing different types of distractions. To be more specific, an on average better performance means the following: the Distraction model is better at modelling data A than model B and at minimum equal to model A. The symmetric case also being true.

With this knowledge a conjecture can be setup with the following null and alternate hypotheses:

H0: The Distraction model is equal or worse at modelling distractions than the existing TCI models.
H1: The Distraction model is better at modelling distractions than the existing TCI models.

6.1.2. Distractions and Data

As mentioned above to validate the model at least two different types of distractions need to be considered. Furthermore as mentioned in chapter 4.1 Distraction Characteristic Level the Distraction model is built upon low-level characteristics, so when determining if the distractions are different it means that the lower level characteristics of the distractions are different.
Normally once the distractions have been chosen, corresponding datasets need to be found. That said since data is the limiting factor in this thesis and topic in general, the order of the distraction selection process will be reversed. Instead two datasets with different distractions will be chosen. Those datasets are then analysed and if their distractions are different enough they will be accepted for the validation test.

The selected datasets are the i4Driving motorway data and CARRS-Q data. The distractions in these two datasets are a continuous mental distraction and a spontaneous phone call. The main low-level attributes of the mental distraction are as follows, it is an internal distraction, a voluntary distraction and mainly via a visual and mental sensory modality. On the other hand the low-level attributes of the phone call are: internal distraction, an involuntary distraction and auditory sensory modality. These two distractions have different low levels characteristics and are thus suitable for the validation test. More details regarding these datasets and their corresponding experiments can be found in the following sections.

The i4Driving data is from the European i4Driving project which has the goal to establish a realistic road safety baseline for the virtual assessment of road safety systems. This project has designed and performed two different experiments with each having multiple scenarios in order to gather data for a multitude of road safety systems. In this validation test only a small subset of recorded data is used, more specifically the University of Naples Federico II (UNINA) highway experiment data scenario 4 and 5 day measurements. Scenario 4 is part of the main phase of the experiment, it uses simulated traffic to emulate the surrounding traffic on an ordinary highway. This is identical for scenario 5 except that the drivers were asked to perform an additional NDRT whenever they felt it was possible. The NDRT used in this experiment is a variation of the Rotated Figures Task from Stanton et al. (1997) which is mainly considered a mental task. (Olstam et al., 2023)
In total this experiment had 32 participants which results in 32 trajectories per scenario. Of these 32 participants, 16 were male and 16 were female. Furthermore 19 were under the age of 31 and 13 were over the age of 31. Lastly 15 of the drivers considered themselves inexperienced and 17 experienced. This means that the total dataset covers a wide variety of people and should properly represent drivers as a whole without being biased towards any particular group, except for cultural driving norms since all participants were from Italy.
To give a better idea of what these trajectories look like two examples of scenario 4 trajectories are given below in figures 6.1 and 6.2. The first is a relatively average trajectory from the dataset and the other is an unusual one where a new leader suddenly cuts in front of the ego vehicle.

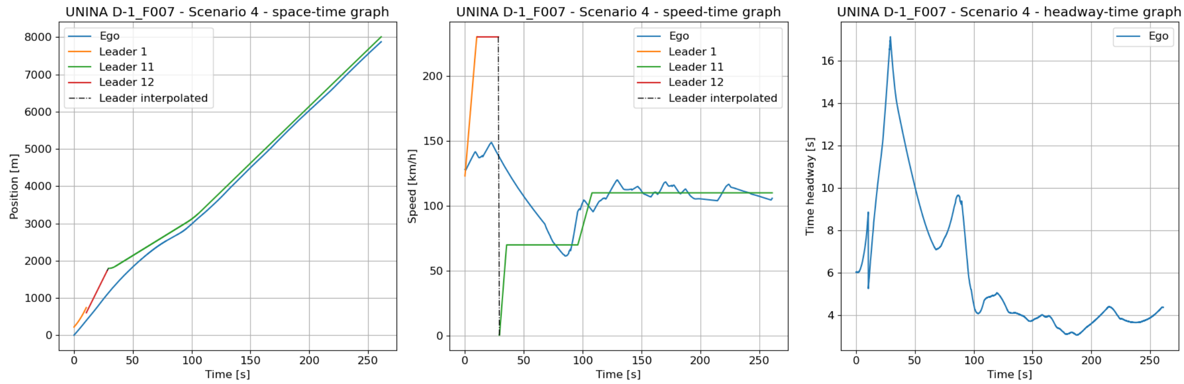

Figure 6.1: UNINA scenario 4 ordinary trajectory.

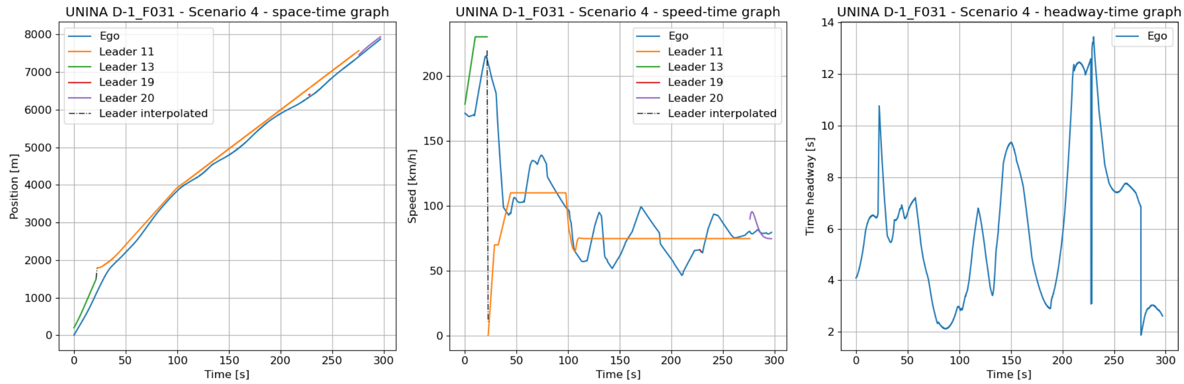

Figure 6.2: UNINA scenario 4 unusual trajectory.

The second dataset is the CARRS-Q dataset. The original dataset was made to accommodate for multiple papers but its primary focus is on the interaction between cars and pedestrians. This simulator experiment was designed and performed by the Centre for Accident Research and Road Safety-Queensland (CARRS-Q). This simulation is repeated multiple times for each participant but under slightly different circumstances each time. The first time, scenario one, is without any distraction, the second scenario is with a handsfree call halfway through and the third scenario is with a handheld call half way through the simulation. For this thesis only a short part, roughly 30 seconds, near the beginning of scenario 1 and 3 will be used, namely the car-following section where the phone call starts.
The simulator experiment has a total of 32 participants. The original paper and data mention that the distribution of gender, culture and other social factors is representative of the local populations in Queensland but doesn't mention the exact ratios or the factors of each participant. The only factor which isn't representative of the local driving population is the age and driving experience of the participants. All participants in the experiment are young drivers between the ages of 18 and 26 years old. Furthermore most participants consider themselves inexperienced compared to other drivers.
To give a better idea of how this data looks, one trajectory from scenario 1 and scenario 3 is given below in figure 6.3 and 6.4.

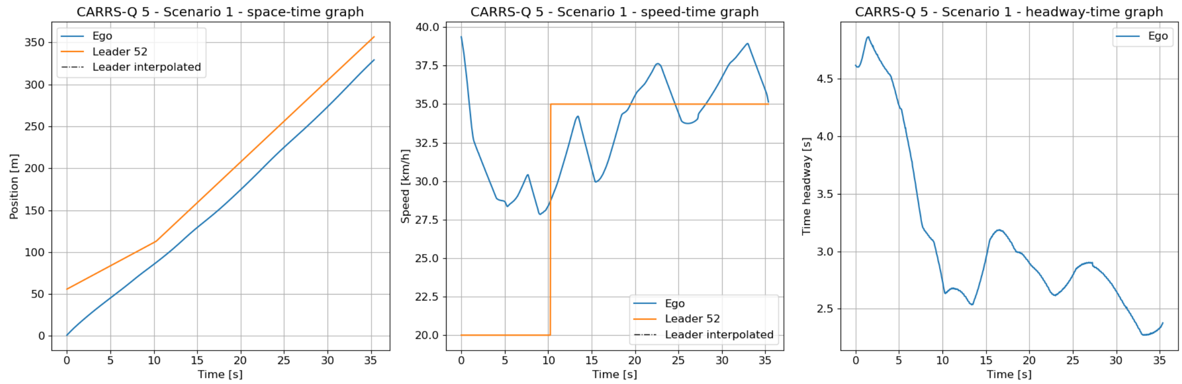

Figure 6.3: CARRS-Q scenario 1 ordinary trajectory.

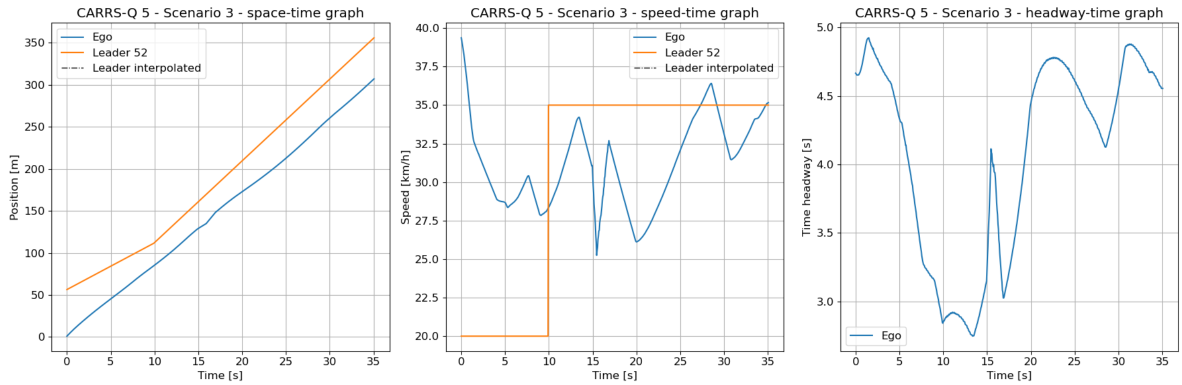

Figure 6.4: CARRS-Q scenario 3 ordinary trajectory.

Lastly to summarise the details of each dataset the characteristics and social factors are compiled in table 6.1.

	UNINA	CARRS-Q
Trajectory duration	~120s	20-30s
Distraction modality	Visual & mental	Auditory
Distraction duration	Full simulation	Half simulation
Spontaneous distraction	No	Yes
Participant age	19 <31 y/o & 13 >31 y/o	18-26 y/o
Participant gender	16M & 16F	equal
Participant experience	15 inexperienced & 17 experienced	mostly inexperienced
Participant culture	Italian	Australian

Table 6.1: Summary of distraction and social characteristics of used datasets.

6.1.3. Comparison with TCI Models

Having acquired two data-sets each with a specific distractions, two TCI models need to be chosen which are capable of representing these two types of distractions. Given that the distractions are, one, a continuous constant distraction and, two, a spontaneous variable distraction. To represent the UNINA data, the continuous constant mental distraction, a good model would be the Multi-scale framework. This framework was made specifically for mental distractions and should have a good performance on this dataset. For the CARRS-Q data, the spontaneous variable auditory distraction, the TDIDM model should perform well. This model was specifically built and trained for this dataset and has a good performance with it. The exact formulas for these models can be found in appendix C Model Formulas.

For more details about these two models see 3.2 TCI Model Implementation and 5 TCI Multi-criteria Analysis

The implementation of these two models is fairly straightforward. The TDIDM model is a complete model which needs no further modification, it only needs to be calibrated on the two datasets. The Multi-scale framework does need some work as the different modules for the task demand need to be chosen. To start, the task demand modules for the driving task are needed. The modules used for this are the ones presented in the original paper of the Multi-scale model by Van Lint and Calvert (2018), see appendix C.3 Multi-Scale. Additionally, a module for the task demand caused by distractions needs to be setup. For the scenarios in the datasets which don't have a distraction this task demand is set to zero. The task demand of the distraction in the UNINA distraction scenarios is set to be a continuous constant as this matches the experiment setup, see equation 6.1. The value of this constant will be determined during the calibration process. As for the task demand of the distraction in the CARRS-Q data, it is set as a constant which starts at the given start time of the phone call in the simulation, see equation 6.2. Again, this constant will be determined during the calibration.

$$TD_d(t) = TD_d \tag{6.1}$$

$$TD_d(t) = \begin{cases} TD_d & t > t_{start,\, i} \\ 0 & else \end{cases} \tag{6.2}$$

with $TD_d(t)$ as distraction task demand variable, TD_d as free distraction task demand parameter, t as time, $t_{start,\, i}$ as phone call start time and subscript i as driver specificity.

Other than the Distraction model and the two models which represent the distraction types a fourth model is used for reference purposes during this thesis. This will be a variant of the Multi-scale framework whose goal is to, as accurately as possible, estimate the task demand of a driver regardless of the known information from the experiment. This model isn't taken into account during the assessment or validation and only serves as a point of reference to determine how well a model could perform. The main difference between this model and the other Multi-scale model is how the task demand for distractions is setup. The task demand doesn't follow the known information about the different datasets or scenarios; instead it divides the task demand over time into five equal length segments, each with their own value, see equation 6.3. This allows for a more precise estimation of the level of distraction which the driver experiences during simulation. This model will be referred to as the Multi-scale Five model.

$$TD_d(t) = \begin{cases} TD_{d,\, 1} & t <= \frac{1}{5} t_{sim} \\ TD_{d,\, 2} & t > \frac{1}{5} t_{sim} \wedge t <= \frac{2}{5} t_{sim} \\ TD_{d,\, 3} & t > \frac{2}{5} t_{sim} \wedge t <= \frac{3}{5} t_{sim} \\ TD_{d,\, 4} & t > \frac{3}{5} t_{sim} \wedge t <= \frac{4}{5} t_{sim} \\ TD_{d,\, 5} & t > \frac{4}{5} t_{sim} \end{cases} \tag{6.3}$$

with $TD_d(t)$ as distraction task demand variable, $TD_{d,\, 1-5}$ as free distraction task demand parameters, t as time and t_{sim} total simulation time.

6.1.4. Assessment Criteria

In order to determine if the Distraction model performs better than the other models a set of assessment criteria needs to be setup. Since the Distraction model is a car-following model the most important part in the assessment is the evaluation of the trajectory.

To measure how good a trajectory is we compare it to the real measured trajectory from the original data. The closer the simulated trajectory is to the measured trajectory the better. To do this the symmetric mean absolute percentage error (SMAPE) of the spacing, speed and time headway is calculated for each simulated trajectory. An explanation for why the SMAPE is used instead of other KPIs is given in 6.3 Model Calibration. Ideally, to conclude that model A is better than model B, the three

SMAPE's of model A should be statistically lower than model B. That said the minimum needed to conclude that a model is better than another would be for one of the SMAPE's to be statistically lower than the other while the other two are statistically equivalent. In the case that at least one of the SMAPE's is statistically lower and at least one is statistically higher then the results are inconclusive and no model can be declared as superior.

To determine if the values are statistically different the SMAPE values need to be calculated for a set of trajectories and then evaluated with a t-test. In this case the appropriate *t*-test is the Welch's unequal variances *t*-test since the two models provide two independent results with different variances (Welch, 1947). Once the t value, see (6.5), is calculated it needs to be converted to a p value with a student's *t*-distribution cumulative density function (CDF), see (6.7), (Bevans, 2020; Wikipedia contributors, 2024, 2025). For this we need two more points of information; first is it a one- or two-tailed test and second what are the degrees of freedom (ν). Since we want to know if one model is more accurate and not just different from another model we use a one-tailed test and the degrees of freedom for the Welch's *t*-test can be seen in (6.6) (Wikipedia contributors, 2024).

$$s_{\bar{X}_i} = \frac{s_i}{\sqrt{N_i}} \tag{6.4}$$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_{\bar{X}_1}^2 + s_{\bar{X}_2}^2}} \tag{6.5}$$

$$\nu \quad \approx \quad \frac{\left(s_{\bar{X}_1}^2 \; + \; s_{\bar{X}_2}^2 \right)^2}{\nu_1^{-1} s_{\bar{X}_1}^4 \; + \; \nu_2^{-1} s_{\bar{X}_2}^4} \tag{6.6}$$

$$CDF \;=\; \frac{1}{2} + t\, \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\, \nu}\;\Gamma\left(\frac{\nu}{2}\right)}\; {}_2F_1\left(\frac{1}{2}, \frac{\nu+1}{2} \,;\, \frac{3}{2} \,;\, -\frac{t^2}{\nu} \right) \tag{6.7}$$

6.1.5. Validation Tests

To test the hypothesis setup in 6.1.1 Proof of Concept for Validation two different approaches are taken. These approaches use the data, models and assessment methods explained above. The first approach tests how well the models can represent a singular trajectory and the second approach tests how well they perform as calibrated models.

The singular trajectory test referred to as individual trajectory calibration consists of calibrating each model to all of the individual trajectories in the two datasets. The optimized results, parameters and KPIs, of each calibration is saved and analysed. The spread of performance for each model gives insight into how accurate and consistent the model is.

The second test, the calibrated model test, relies on the results of the individual trajectory calibrations. To perform this test the data for each scenario is randomly split into two subsets, the calibration subset and validation subset. The calibration subset consists of 20 trajectories and the validation subset consists of 12 trajectories. Using the parameter results found in the individual trajectory calibrations the mean parameters of the calibration subset are calculated. These mean parameters then serve as basis for the calibrated model. The performance of the calibrated model is measured with the validation set. Since the split datasets are quite small the test is run three times with three different data splits, called seeds, to ensure that the results aren't biased to any particular individual trajectory. The calibrated test provides different insights into the model. It shows how well the model performs on unknown data and by comparing it with the individual trajectory calibration results it shows how much the model tends to overfit.
One particularity of note is that the Multi-scale Five model is excluded from this test. This model is designed to overfit the individual trajectories and makes assumptions reinforcing this fact. Due to these assumptions it is meaningless to make a calibrated version of this model meant for multiple unknown trajectories.

6.2. Distraction Model Implementation

Knowing the datasets and scenarios the final form of the Distraction model can be setup. The main Distraction model logic is based on the Multi-scale model whose formulas can be found in appendix C.3 Multi-Scale. The majority of the modifications made to the model concern the distraction task demand ($TD_d(t)$), figure 6.6 provides an overview of which parts of the Multi-scale model are modified. The distractions are modelled using the steps and definitions setup in 4.2 Distraction Framework. Multiple variants of the Distraction model were tested and only the variant yielding the best results was kept and is explained below. Each of the tried variants are also listed below.

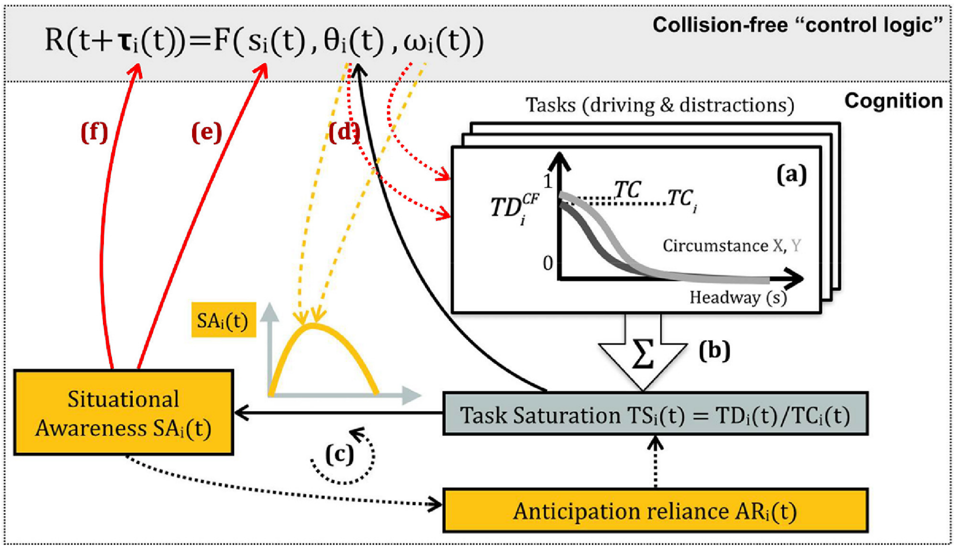

Figure 6.5: Multi-scale model relationships modified by the distraction framework in red. Figure source: (Calvert et al., 2020) (Modified)

The first step in the Distraction model is the distraction trigger (DT_i). Since the trigger depends heavily on the type of distraction the two datasets need a dataset specific trigger. This is the only formula in the Distraction model which is dataset specific. The two formulas for the DT_i can be seen in equations 6.8 and 6.9 being the triggers for the UNINA and CARRS-Q dataset respectively. These equations are located at arrows (d) in the logical flow diagram in figure 6.5.

$$DT_i(x_i, t) = \left(\sum_{t_g = t - t_{linger}}^{t} gaze_{dis}(t_g) \right) > 0 \tag{6.8}$$

$$DT_i(x_i, t) = t > t_{start,\ i} \tag{6.9}$$

with t_{linger} as distraction linger time, $gaze_{dis}(t)$ gaze direction distraction and t_{start} as distraction start time. The distraction linger time determines how long a distraction stays active after having looked at a distraction. The gaze direction distraction is a binary variable which tells if the driver is looking at a distraction or not and lastly the distraction start time is the time at which distractions are present or start appearing.

The motivation behind the equations are as follows. The UNINA visual distraction is only applicable when the driver looks at the screen with the question and a short time afterwards while the driver answers the question. The exact manner of incorporating gaze data into the formula is not based on literature since no relevant literature was found using gaze data in car-following models or other operational-level decision making models. The only model which applied gaze data are tactical-level decision making models. The CARRS-Q distraction is dependant on a spontaneous phone call, this is far simpler to model since only the start time of the first phone ring is needed which is provided by the dataset.

These two equations also make use of a couple of constants, namely the linger time and the start time. The linger time is set to 2 seconds, multiple options were tried and 2 second provides the best and most consistent results. The start time is simulation/driver specific, it is provided by the CARSS-Q dataset and is setup during initialisation of the scenario-driver pair.

The second equation is the distraction intensity in equation 6.10. As mentioned before only one equation is needed to cover both datasets. This equation is located at arrows (d) in the logical flow diagram in figure 6.5.

$$DI_i = \begin{cases} TD_{d,\,i} = DS * DDS_i & SP = 0 \ \vee t > t_{start,\,i} + 2s \\ TD_{dsp,\,i} = DS * DDS_{SP,\,i} & SP = 1 \ \wedge t > t_{start,\,i} \end{cases} \tag{6.10}$$

$$TD_{total,\,i} = DI_i + TD_{cf} \tag{6.11}$$

with $TD_{d,\,i}$ as distraction task demand, $TD_{dsp,\,i}$ for spontaneous distraction task demand, free parameter $DDS_{SP,\,i}$ as spontaneous driver distraction sensitivity and $t_{start,\,i}$ as distraction start time. The two task demands are shorthand for the distraction intensity that separates normal and spontaneous distractions. The spontaneous driver distraction sensitivity serves the same function as the normal driver distraction sensitivity but specifically for spontaneous distractions and the distraction start time is the time at which distractions are present or start appearing.

The logic behind this equation depends mostly on whether a distraction is spontaneous or not. In case of a normal distraction the distraction intensity is a standard constant rate. If the distraction is spontaneous it gets a modified rate for two second and afterwards it returns to normal distraction rate. The two second duration isn't chosen at random but is instead based on the average time it takes to pickup the phone call found in the data. Furthermore since the distraction strength is a constant and the driver only experience one type of distraction during the simulation it can be simplified into a singular value which for ease of use is called distraction task demand. An example output of this equation can be seen in figure 6.6. The second equation, 6.11, shows how the distraction intensity is implemented in the Multi-scale model. It is simply added to the other task demands to calculate the total task demand. The distraction intensity equation does make use of a couple of constants. These are the distraction strength, which is set to 1 for convenience since only a single distraction is present in the simulations. The two driver distraction sensitivity parameters which are estimated during calibration and are driver specific. The spontaneous factor (SP) which represents if a distraction is spontaneous or not and is tied to the distraction/scenario. Lastly the distraction start time is used in the same manner as in equation 6.9. For UNINA it is simply 0 since the distraction starts right away and for the CARRS-Q dataset it is the start of the phone call.

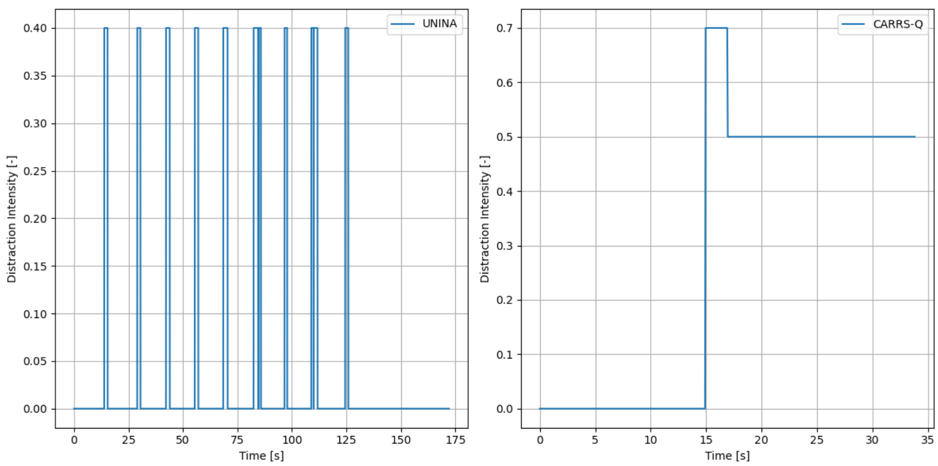

Figure 6.6: A fictive example of the distraction intensity from the UNINA and CARRS-Q datasets.

The last equation is a form of distraction effect. In the standard formulation of the Multi-scale model there are already a few equations which could be considered distraction effects, see appendix C.3 Multi-Scale. The Distraction model adds onto those equations with equation 6.12. This equation is located at arrow (e) in the logical flow diagram in figure 6.5.

$$a_{max,\ i}(t) = (1 - \beta_{a,\ i} * DI) * a_{max} \tag{6.12}$$

with free parameter β_a as maximum acceleration reduction factor and $a_{max,\ i}(t)$ as specific maximum acceleration. The maximum acceleration reduction factor determines by what fraction the acceleration can be reduced when experiencing a high distraction load. Lastly the specific maximum acceleration is a driver and time-dependent version of the maximum acceleration which is used in the final acceleration equation of the Multi-scale model.

This equation was added with some trial and error after analysing the predicted trajectories and comparing them to the real trajectories. One notable reoccurring flaw was that the vehicle accelerations were too high when under a high task demand. This effect is further supported by the correlation between the effort and the absolute acceleration. This correlation has a R-squared of 0.24, while weak it does show that there is a connection between the two. Multiple variants were tried each with their own pros and cons, for details of the variants see below, the best performing variant is the version which is listed above. Unlike the initial definition of the formula it directly uses the distraction intensity instead of the total task demand or situational awareness.
This equation uses a single new constant the maximum acceleration reduction factor. It is a driver specific parameter that is determined during the calibration process.

6.2.1. Distraction Model Variants

As mentioned above the final equations are the result from trial and error. In order to keep track of which variants improved the results and which variants worsened the results a short overview of each tried variant is listed below. The general workflow for formulating the equations is as follows. First find potential areas of improvement in the graphs of the trajectories, second study literature and correlations in the data for existing solutions to the found weaknesses and third implement and adapt solution with trial and error to find the best variant.

The first equation which had different variants is the UNINA distraction trigger. Since the UNINA distraction is a combination of a visual and mental modality one idea was to use gaze data for the trigger. After all a person should only be distracted from the moment they look at the distraction. As mentioned before, literature contains no existing implementations of gaze data in an operational-level model. The few models which did use gaze data were on a tactical level and offered little to no insight on how to apply this on an operational level (Kircher et al., 2013; Phan et al., 2016; Radhakrishnan et al., 2023). Instead the effort measurement from the UNINA dataset were used as a guideline. The idea is that high effort from the drivers means a high task demand (Irvine et al., 2023; Tang et al., 2023) which is likely caused by the distraction. With the goal of trying to match most of the effort peaks with the gaze data some trial and error tests were performed. These trial and error tests operated with two criteria, first when does someone count as distracted and second how long are they distracted. For the first criteria a couple of different options were tried, first a driver is distracted when they look at the question screen, second a driver is distracted when they look inside the vehicle, third a driver is distracted when they don't look at the road and last a driver is distracted when they don't look at the road ahead. Of these cases the best results were obtained by the first option of looking at the question screen. The second criteria also had a couple of alternatives. First the distraction stops when the driver look away, second the distraction lingers for 1/2/3/5 seconds after looking away and last the distraction lingers for 1/2/3/5 seconds after looking away but decreases linearly to zero. Out of these a 2 second lingering constant distraction offered the best performance. Its performance is equivalent to 1 second lingering case while the lingering cases above 2 seconds deteriorate the performance. Additionally having a linearly decreasing distraction level seems to have little to no effect on the performance.
Contrary to the UNINA distraction trigger, the CARRS-Q distraction trigger has no variants. The dataset is quite clear about how the distraction functions and when it starts.

In similar fashion to the UNINA distraction trigger equation, the distraction intensity equation also had different variants. The purpose of this second equation is to determine the strength of the distraction.

In both the UNINA and CARRS-Q dataset the strength of the distraction remained constant during the simulation but participants were asked to only complete the distraction when they felt comfortable engaging with them. This means that UNINA participants only had to look and answer questions when they wanted and CARRS-Q participants could choose to delay their responses to the phone conversation. The variants for the distraction intensity tried to capture this phenomenon. The idea was as follows, if a driver is allowed to limit his engagement with the distraction, thereby limiting the intensity of the distraction, he would likely use this option when his total task demand exceeds his capacity. This concept is quite similar to the anticipation reliance found in the Multi-scale model but it uses a different reasoning basis (Calvert et al., 2020). The implementation of this concept was as follows. The distraction intensity equation applied an additional reduction factor when the total task demand exceeded the task capacity, that said this reduction factor wasn't applied to spontaneous distractions. This variant provided a minor insignificant improvement but it nearly doubled the standard variation of the performance. For this reason it was decided to not use this variant despite the minor improvement. A different variant which attempted to model this effect used the pupil diameters. The effort in the UNINA data is calculated from a couple of different parameters, one of which is the pupil diameter. Since the i4Driving papers assume a correlation between task demand and effort (Irvine et al., 2023; Tang et al., 2023) it should also be possible to tie the pupil diameter to the distraction intensity. The relation between the two is assumed to be the following; if a person experiences a high task demand, likely due to a distraction, they will have a wider pupil diameter. Logically if this relationship is inverted it should also be possible to calculate the task demand from the pupil diameter. With this in mind multiple attempts were made to modify the distraction intensity. The first attempt directly applied the pupil diameter as scaling factor, the second attempt tried adding the pupil diameter as scaling factor when it exceeded a threshold and the last attempt tried the opposite of reducing the intensity when the pupil diameter stay under a threshold. None of these attempts had a positive effect on the performance and the third attempt had a negative impact on performance so they were not kept in the distraction intensity equation.

The last equation, the distraction effects, also had different variants. More specifically the new distraction effect equation, the modified maximum acceleration, was finalized after some trial and error. The overall goal of this equation is to reduce the acceleration when appropriate. This is similar to the other parameter reduction equations seen in appendix C.3 Multi-Scale, so similar attempts were made. The first variant followed the same formulation with the maximum reduction factor multiplied by the situational awareness error, the second variant tried using the task saturation error, the third used the total task saturation directly and the last variant of this formulation used the distraction intensity. Other variants included a quadratic reduction error and an error term which was additive instead of multiplicative. Of these variants the multiplicative linear distraction intensity reduction factor provided the best results. While most variables had similar results the distraction intensity was the only variable providing additional value. Furthermore the quadratic and additive reduction factors provided no significant changes compared to the linear multiplicative reduction factor.

6.2.2. Distraction Model Verification

With the mathematical form of the Distraction model finalised it should now be tested to see if it behaves as defined in theory. This test is the verification test which is composed of multiple steps. To verify if the model behaves correctly all the new parameters are tested one by one. The changes in behaviour from the model are then observed and compared to the a unmodified version of the model. If the changes match expectation then the model passes the verification step.

The chosen default parameters of the Distraction model can be seen in table 6.2. The used values can be considered as the common standard values for these parameters. The altered values are also given in the table, each of these altered values will be tested one by one while keeping other values unmodified. The results of this test can be found in 7.1 Distraction Model Verification. It should also be noted that the leader vehicle in the test is a car which uses the IDM model with the same parameters except for a slightly lower desired speed to ensure proper car-following behaviour.

Parameter	Value	Altered value	Unit
b_{max}	8	-	m/s^2
a_{max}	4	-	m/s^2
b_{comf}	3	-	m/s^2
v_0	30	-	m/s
s_0	8	-	m
T	1.2	-	s
τ	0.5	-	s
δ_{sign}	0	1	$-$
TD_d	0.4	0.8	$-$
TD_{dsp}	0.7	0.3	$-$
β_a	0.6	0.2	$-$

Table 6.2: Distraction model parameters for verification.

6.3. Model Calibration

In order to get proper results from the models, the models need to be calibrated on the trajectories of the test data. This is often achieved using optimisation algorithms. This thesis will make use of Genetic Algorithm (GA) to find the ideal parameters to fit the model to the real trajectories. Genetic Algorithm are commonly used in car-following models since they tend to avoid local minima and use stochastic methods to find global minima (Kesting & Treiber, 2008; Punzo et al., 2012). The main downside of using GA to calibrate the models is that it is computationally intensive, though there are tools available which can mitigate this downside by implementing good code optimisations and multi-threading computation.

When using GA the main difficulty and most important step is deciding the objective function, also known as fitness function, of the optimisation. Using this objective function the algorithm is able to tell how well the model performed. Since the goal of the calibration is replicating the real trajectory a similar approach to the assessment criteria is taken. In the assessment criteria of the validation experiment, the ”goodness” of the trajectory is measured with the symmetric mean absolute percentage error (SMAPE) of the position, speed and time headway. Ideally all three values could be given to the GA as objective function but the used GA software, explained below, only supports a singular value as objective function. While the objective function could be some arbitrary combination of these KPIs it was opted to use a singular KPI as objective value since an arbitrary combination can't be interpreted as a real world performance. Out of the three performance parameters, the SMAPE of the headway provides the most information since this not only provides information on the position but also on speed and consequently the error in both get minimized (Kesting & Treiber, 2008).

The SMAPE is the mean absolute percentage error, first introduced by Armstrong (1985) and later further refined by Flores (1986). This KPI isn't the most commonly used error measurement metric but it provides a few interesting benefits over the more traditional metrics like the root mean squared error (RMSE) or root mean squared percentage error (RMSPE). There are two benefits with this metric. First it supports zero values for either the observed or predicted values. Secondly it limits the error to be between 0 and 100%, this means that a few non-zero extremely small or extremely large observed values won't cause a major shift in the mean error. Both of these characteristics are relevant given our datasets since speeds can get close to or reach zero while the headway can reach near infinite in some of the observed trajectories. The major drawback of this metric is that it has little to no real world interpretation unlike the RMSE or RMSPE which can be interpreted as a real physical parameter. The formula for the SMAPE is as follows:

$$SMAPE_h = \frac{1}{N}\sum_{n=1}^{N}\frac{||h_{sim} - h_{real}||}{||h_{sim}|| + ||h_{real}||} \tag{6.13}$$

6.3.1. Optimisation Parameters

Other than the objective function there are a few other parameters to configure before a model can be calibrated. These are the population parameters and the gene parameters. The first controls how long and how extensively the GA will search for solutions while the second helps scope the search.

In this thesis the python package *PyGAD* is used as a tool to implement the Genetic Algorithm; PyGAD provides a few parameters to control the population, namely the $num\_generations$, $num\_parents\_mating$, sol_per_pop. For this thesis these are set at 250, 4 and 8 respectively. Depending on the complexity of the model and the number of parameters that need to be calibrated these parameters can be increased but this provides a baseline when first calibrating a model. In addition to the total number of generations limit, an additional constraint is implemented which stops the optimisations after 50 generations if there is no improvement.

The genes are the representation of the model parameters which need to be optimized. The number of genes usually equals the number of model parameters. As for gene parameters there are two in total. First is the $mutation\_percent\_genes$, set at 25%, which controls how often a gene can change between generations. The other parameter is the gene bounds or gene space, this controls the limits for how far the algorithm will search to find an optimal solution. The gene bounds are model dependent and are given below for each model.

Table 6.3 lists all the common shared parameters. These values are based from a study by Papathanasopoulou and Antoniou (2015), physical limits, extrapolation from the data and intuition. The b_{max} and a_{max} have been fixed as these are physical characteristics of the vehicle, the values of these parameters are deduced from the data. The bounds defined in table 6.4 are copies of the calibration bounds presented by Saifuzzaman et al. (2015). These bounds correspond to the range where these parameters still have a feasible real world representation. The parameter calibration bounds in table 6.5 are set to the defined theoretical minimum and maximum values for each corresponding parameter.

Parameter	Min	Max	Step	Unit
${b_{max}}^*$	8	8	-	m/s^2
${a_{max}}^*$	4	4	-	m/s^2
b_{comf}	0.5	5	-	m/s^2
v_0	10	50	-	m/s
s_0	1	15	-	m
T	0.5	5	-	s
τ	0.1	2	-	s

Table 6.3: Common model free parameter bounds for GA optimisation. *Fixed values, based on test data.

Parameter	Min	Max	Step	Unit
δ	-5	1	-	–
γ	0.5	5	-	-
φ	0.1	2	-	s

Table 6.4: TDIDM model free parameter bounds for GA optimisation (Saifuzzaman et al., 2015).

Parameter	Min	Max	Step	Unit
δ_{sign}	0	1	1	-
TD_d	0	1	-	-
TD_{dsp}	0	1	-	-
β_a	0	1	-	-

Table 6.5: Multi-scale and Distraction model free parameter bounds for GA optimisation.

6.3.2. Synthetic Data Test

To test if the GA is correctly configured it is tested with synthetic data. This means that a trajectory is generated with a validated CF model using known parameters. This trajectory is then used as the reference trajectory during the optimisation process of the same model but with unknown parameters. If the parameters of the optimized model are close to the original parameters then the test is successful and the GA can be used for the validation experiment.

For this test the Multi-scale model will be used. Using this model three cars will be simulated which follow one another, these will be called leader 1, leader 2 and ego. This test is run for three different scenarios. All scenarios use the same model parameters and a time step of 0.05 seconds but each has a different total duration. In the first scenario, all cars experience a strong distraction halfway through the scenario until the end of the scenario. This scenario lasts 60 seconds. In the second scenario the leader slows down in the middle of the scenario and speeds back up shortly after; this scenario lasts 100 seconds. The last scenario is a combination of both previous scenarios, it lasts 250 seconds, it has two slowdowns, one in the first half of the scenario and one in the second half while the driver is also distracted. The goal of each scenario will be to replicate the trajectory of the last car, the ego. The exact parameters of the model of the optimisation can be found below:

Parameter	Synthetic
b_{max}* $[m/s^2]$	8
a_{max}* $[m/s^2$	4
b_{comf} $[m/s^2]$	3
v_0 $[m/s]$	35
s_0 $[m]$	8
T $[s]$	1.2
τ $[s]$	0.5
δ_{sign} [-]	0
TD_d [-]	0.7

Table 6.6: Synthetic Multi-scale model optimisation test parameters. *Fixed parameters

7
Results

This chapter presents the results from the experiments explained in 6 Validation Methodology. These results will be presented in the same order as they are explained in chapter 6. This means that the results regarding the individual trajectory calibrations are given first and the results of the calibrated model are second.

7.1. Distraction Model Verification

To ensure that the distraction model functions as theorized, a verification test is carried out. The results of this test can be seen below.

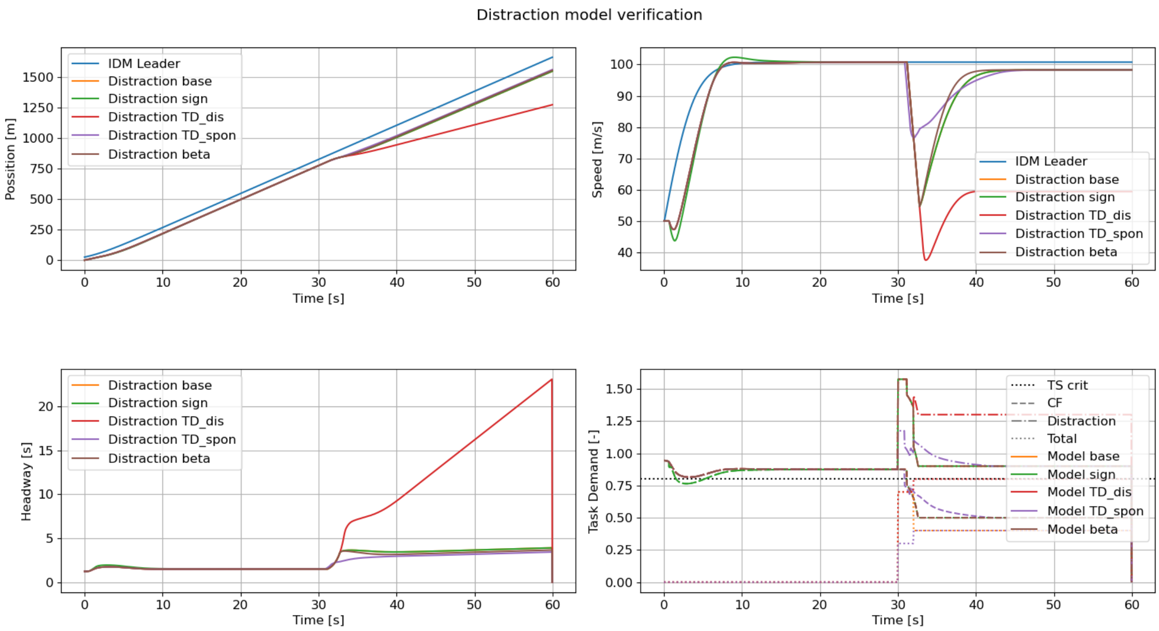

Figure 7.1: Trajectories of Distraction model and verification variants.

Overall the results of the verification test show that the model behaves as expected. The following section will elaborate on each modified parameter and will explain the observed results.
The first parameter which was tested is the δ_{sign}, which determines if a driver is over- or under-estimating when overloaded. In the base parameters the driver is overestimating his speed and spacing differences with the leader and in the modified test they are underestimating. Based on that it would be expected that the altered drivers keep a longer headway than drivers at lower speeds compared to the baseline case. This is exactly what is observed in the results. Additionally two other deviations stand

out, first the driver reaches a higher top speed when catching up in the first section due to the larger spacing and second the driver has a lower CF task demand than all other drivers in the start as they are more conservative.
The second modified parameter is the TD_d, which control the task demand of normal distractions. The altered value of this parameter is much higher than the default value. This should cause the total task demand to increase during the second half of the simulation and since it will cross the critical task saturation level by a significant margin it will also significantly reduce the desired speed. This effect can also be observed in the results, the speed and task demand graphs show the decrease in speed and increase in task demand respectively.
The third parameter is the TD_{dsp}, which controls the task demand of spontaneous distractions. Compared to the base value the altered value is much lower, due to this there should be a much milder decrease in speed at the moment of the distraction. As expected this effect is also reflected in the results. In addition to this effect another change in the results is that recovery from this speed loss is much more gradual than in the baseline model. Furthermore this gradual increase in speed also causes the car-following task demand to decease to the minimum value in a much more gradual manner. Both of which can be explained by the non-linear nature of the underlying IDM model.
The last parameter is the β_a, which applies a reduction in maximum acceleration when distracted. Since this reduction factor is decreased in the altered case it would be expected that the car accelerates faster after the speed drop seen in the velocity graph. This is indeed what happens, the vehicle accelerates faster and reaches their desired speed a few second earlier. At first the effect looks less pronounced than was expected but that is due to the scaling of the graph which squishes the acceleration period making it look similar to the baseline case.

7.2. Synthetic Data Test

The synthetic data test is performed to test whether the genetic algorithm performs as expected. An overview of the results of this test can be seen in table 7.1. A more in-depth analysis of the results is provided in the following paragraphs.

Parameter	Synthetic	Scenario distraction	Scenario slowdown	Scenario complex
b_{max}* $[m/s^2]$	8	-	-	-
a_{max}* $[m/s^2$	4	-	-	-
b_{comf} $[m/s^2]$	3	2.94 (-2.0%)	3.58 (+19.3%)	2.99 (-0.3%)
v_0 $[m/s]$	35	34.99 (-0.0%)	35.06 (+0.2%)	34.99 (-0.0%)
s_0 $[m]$	8	8.25 (+3.1%)	1.71 (-78.6%)	11.17 (+39.6%)
T $[s]$	1.2	1.10 (-8.3%)	1.48 (+23.3%)	1.06 (-11.7%)
τ $[s]$	0.5	0.51 (+2.0%)	0.48 (-4.0%)	0.50 (+0.0%)
δ_{sign} [-]	0	0.5 (0.0%)	0 (0.0%)	0 (0.0%)
TD_d [-]	0.7	0.70 (0.0%)	-**	0.70 (0.0%)

Table 7.1: Synthetic Multi-scale model optimisation test parameters. *Fixed parameters. **Not present in scenario.

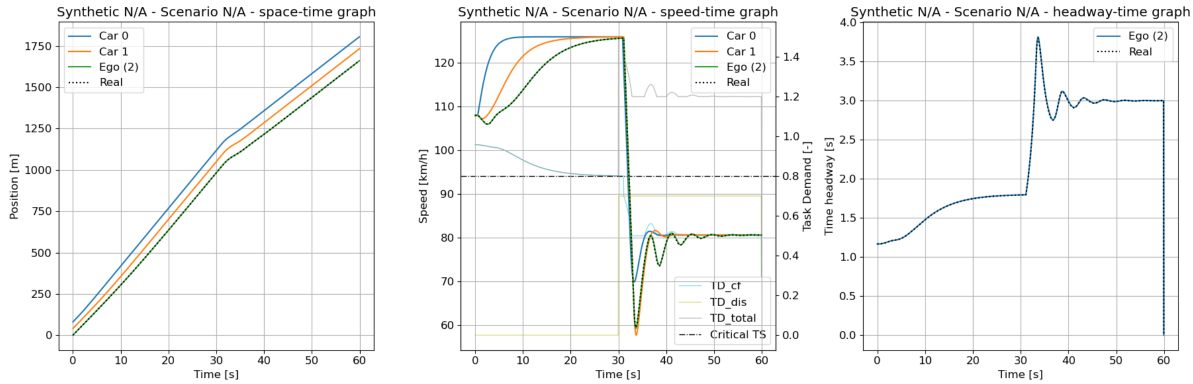

Figure 7.2: Synthetic Multi-scale model optimisation trajectories for distraction scenario.

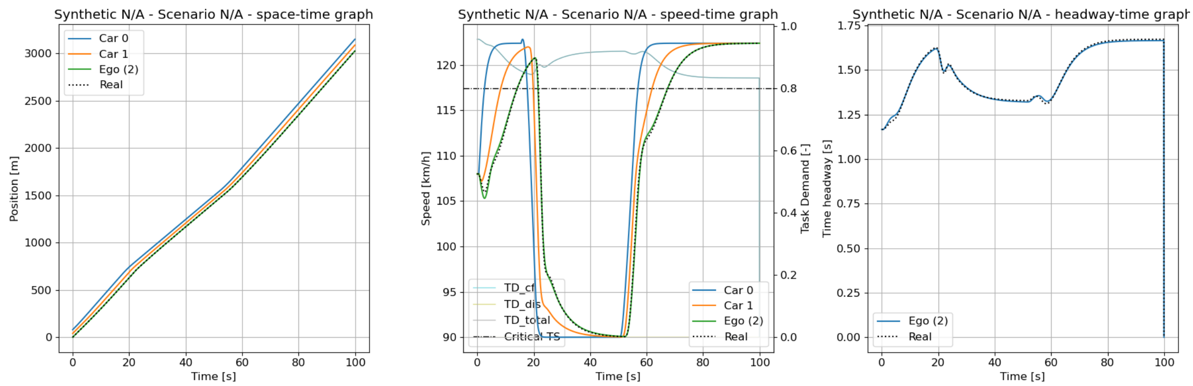

Figure 7.3: Synthetic Multi-scale model optimisation trajectories for slowdown scenario.

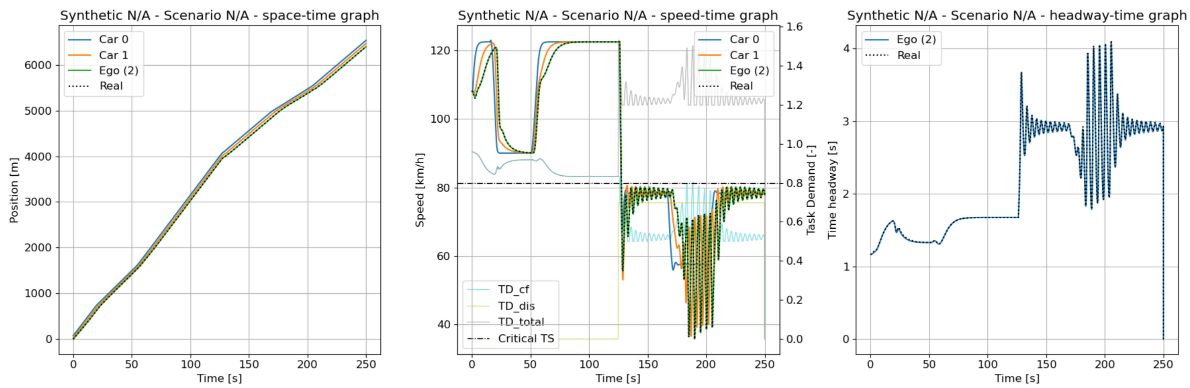

Figure 7.4: Synthetic Multi-scale model optimisation trajectories for complex scenario.

Figures 7.2, 7.3 and 7.4 show the trajectories of the original simulated synthetic vehicles and the found calibrated vehicles. In these graphs there are a few minor discrepancies between the real and found trajectories but overall they are quite close.

This seems to match the various KPI, the best results are from the position data, with a RMSE of 0.06, 0.21 and 0.16 metres. The speed data resulted in slightly worse performance with an RMSE of 0.02, 0.06 and 0.04 m/s. Lastly the headway had the worst performance despite it being the objective value, it has an RMSE of 0.004, 0.007 and 0.011 and a SMAPE of 0.05%, 0.2% and 0.15%. Overall these KPIs are excellent with little room for improvement. Additionally in terms of fitness there is no real noticeable difference between the three scenarios with all scenarios nearing their optima. An important detail to note is that each optimisation was run multiple times with different seeds, the presented results are the ones with the highest fitness values.

When looking at the found parameters, table 6.6, there are two major discrepancies. These are in the b_{comf}, s_0 and T. The easiest to explain is the s_0, the desired headway at standstill, this parameter plays a negligible role at high speeds and becomes much more important at lower speeds. The simulation mainly used high speeds in the expected range of a motorway. Due to this lack of low speed driving the algorithm probably had difficulties finding the correct value for s_0. This discrepancy in s_0 also seems to be correlated with the discrepancy in T. Both parameters influence the desired headway of the driver, and in the two scenarios with discrepancies in one of the parameters the other parameters is also faulty but in the opposite direction. So it's likely that the headway error caused by one parameter gets cancelled out by the headway error of the other parameter. The b_{comf}, comfortable breaking speed, in the slowdown scenario is another parameter whose error can be explained when looking at the speed trajectory. Normally the b_{comf} is correlated to the more gradual decelerations the vehicle would make when there is little to no external influences and the b_{max} is the maximum deceleration caused by external influences. In the speed trajectory it is clear that the car has one big deceleration moment, this deceleration is likely close to b_{max} and there are no other decelerations. The error in b_{comf} is likely due to the lack of data regarding these more gentle decelerations.
Lastly the remaining parameters are relatively close to their synthetic counterparts. This is also true for the TD_d and τ. This is a positive outcome since these two parameters are most commonly used to indicate how distracted a driver is. Since distractions are the primary objective of the validation experiment it is good that the estimation of the distraction parameters are fairly accurate.

7.3. Individual Trajectory Calibration Results

This section covers the results of the individual trajectory calibrations explained in 6.1.5 Validation Tests.

7.3.1. UNINA Data

Following the assessment methodology and KPI given in 6.3 Model Calibration the following results are found and summarized in table 7.2, 7.3, 7.4, 7.5. A more detailed analysis of the results in the tables can be found under each table.

Model	Mean	Median	Min	Max	Std
IDM					
$SMAPE_x$ [m]	98.6%	99.2%	91.4%	99.7%	1.6%
$SMAPE_v$ [m/s]	95.5%	96.1%	87.3%	97.7%	2.4%
$SMAPE_h$ [s]	84.9%	87.2%	53.2%	96.3%	9.5%
TDIDM					
$SMAPE_x$ [m]	98.8%	99.0%	94.6%	99.8%	0.9%
$SMAPE_v$ [m/s]	95.3%	95.6%	87.6%	97.9%	1.9%
$SMAPE_h$ [s]	85.5%	87.1%	64.9%	96.5%	7.1%
Multi-scale					
$SMAPE_x$ [m]	98.6%	99.2%	90.7%	99.8%	1.9%
$SMAPE_v$ [m/s]	95.3%	96.0%	86.1%	97.9%	2.5%
$SMAPE_h$ [s]	84.6%	87.2%	49.2%	96.4%	10.5%
Multi-scale Five					
$SMAPE_x$ [m]	99.3%	99.4%	98.3%	99.8%	0.4%
$SMAPE_v$ [m/s]	96.1%	96.7%	92.0%	98.1%	1.5%
$SMAPE_h$ [s]	91.0%	91.3%	81.1%	96.8%	4.0%

Table 7.2: Individual trajectory calibration SMAPE results UNINA scenario 4 (no NDRT).

The results of scenario 4, the scenario without a NDRT, show that all the models are nearly equal.

Between the IDM, TDIDM and Multi-scale framework models there is practically no difference in mean or median value in any of the KPI. While there is no difference in mean or median values there is a slight difference in the distribution of the results which is reflected in the standard deviation and outliers. The TDIDM model seems to have a slightly tighter distribution of results, compared to the IDM and Multi-scale model, which can be seen in its lower standard deviation and higher minimum value. That said this is a misleading result caused by a singular outlier from participant D-1_F031. When this participant is excluded from the results for all models the gap in standard deviation and minimum value close quite a bit and become nearly equal around 8% for the std and 60% for the minimum value. As for why participant D-1_F031 performed better in the TDIDM model compared to the other models is difficult to accurately explain without extensive testing. That said it is likely that this result is due to luck during the calibration process and that the genetic algorithm happened to fall in a narrow global minimum. This anomaly could not to be reproduced when running the same GA on the same trajectory again with the TDIDM model.
Beyond the three baseline models there is also the result from the Multi-scale Five model. Since this model doesn't comply to the rules and limitations set in the design setup but instead has access to more information than normally available it is expected that this model performs better. This is indeed the case, the model has a slightly higher mean and median value for the spacing and speed KPI and when looking at the headway KPI the model performs much better than the other three models. This is again the case when looking at the distribution of the results, we notice a slight improvement in speed and spacing and a large improvement when looking at the headway.

Model	Mean	Median	Min	Max	Std
IDM					
$SMAPE_x$ [m]	98.9%	99.0%	96.9%	99.7%	0.7%
$SMAPE_v$ [m/s]	92.8%	93.0%	86.1%	96.1%	2.1%
$SMAPE_h$ [s]	86.7%	87.5%	76.3%	95.4%	4.6%
TDIDM					
$SMAPE_x$ [m]	98.9%	99.0%	97.4%	99.7%	0.6%
$SMAPE_v$ [m/s]	93.0%	93.2%	89.1%	96.9%	1.9%
$SMAPE_h$ [s]	86.0%	87.0%	67.5%	94.4%	5.6%
Multi-scale					
$SMAPE_x$ [m]	99.0%	99.0%	97.4%	99.8%	0.7%
$SMAPE_v$ [m/s]	93.2%	93.3%	88.7%	97.1%	1.9%
$SMAPE_h$ [s]	85.8%	87.3%	66.7%	93.8%	5.9%
Distraction model					
$SMAPE_x$ [m]	98.8%	99.1%	95.4%	99.8%	1.1%
$SMAPE_v$ [m/s]	92.5%	93.3%	81.2%	97.1%	3.1%
$SMAPE_h$ [s]	86.8%	87.9%	75.6%	94.2%	5.1%
Multi-scale Five					
$SMAPE_x$ [m]	99.2%	99.4%	97.7%	99.7%	0.5%
$SMAPE_v$ [m/s]	93.4%	93.4%	89.2%	96.9%	1.6%
$SMAPE_h$ [s]	90.3%	90.1%	80.1%	95.3%	2.7%

Table 7.3: Individual trajectory calibration SMAPE results UNINA scenario 5 (yes NDRT).

When comparing the results of scenario 5 to the results of scenario 4 there are a few differences. The main difference is that the results of the different KPIs in scenario 4 tend to be a bit higher than those in scenario 5. This is in line with what would be expected from these two scenarios. Since scenario 5 contains distractions it would be expected that those trajectories contain a bit more variation and would be harder to model. This is the case for most KPI but not all, a notable exception is the headway KPI of the IDM model which has a much higher mean value in scenario 5 compared to scenario 4.

When looking into the results of scenario 5 we see similar results to the results of scenario 4 with a few differences. Overall the three base models, IDM, TDIDM and Multi-scale model performed quite similar to each other with the exception of the headway KPI of the IDM model. The mean headway KPI

of the IDM model is 0.7% and 0.9% higher than those of the TDIDM and Multi-scale model and the std is a full percent lower than the std of the other two models. At first glance this might look similar to the D-1_F031 participant anomaly in the TDIDM model found in scenario 4 but this doesn't turn out to be the case when looking more in-depth into the individual results. When looking into the individual results there are two points that stand out which support that this increase isn't due to a lucky anomaly. First there are multiple outliers in the TDIDM and Multi-scale model which fall under the minimum headway threshold of the IDM model, so even if you remove one outlier it doesn't improve much the results of the other two models. The second point which supports the IDM model is that the ~0.8% improvement in mean value is noticeable in nearly every individual result. Nearly every individual participant result is between 0.5% and 1% higher than their counterpart in the TDIDM of Multi-scale model.
This result is quite unexpected when looking at the setup of the test, data and models. Scenario 5 is the scenario where there are distractions present that should influence driving behaviour in a negative manner. Given this line of thought it would be expected that the models built to capture additional human factors, TDIDM and Multi-scale, behave better then the basic model they are built upon. But this is not the case, not only are they not better, they aren't even equal to the IDM model, instead they are worse. A more in-depth analysis of why this could be the case is further explored in chapter 8 Discussion & Recommendation.

Other than just the baseline models, scenario 5 also included the Distraction model. The most promising Distraction model variant for this dataset was the Distraction model which made use of the gaze data, whose formulation can be found in 6.2 Distraction Model Implementation. The expectation and objective of this model was to perform better than the other HF baseline models. As can be seen in the table this has been accomplished as it has a similar or better mean value for all KPIs. That said its std for those KPIs are a bit worse for the spacing and speed and better for headway. But as was remarked before the two baseline models, for which the hypothesis was formulated, aren't the ones who perform the best in this scenario. The model which performs the best is the IDM model and compared to the IDM model the Distraction model performs nearly equivalent. The mean values of the various KPIs are the same and the only difference is a minor change in distribution, with the distribution of the Distraction model being slightly worse than the distribution of the IDM model. While it is still concerning that the IDM model is the best performing model for this dataset it is also reassuring that the Distraction model is at least equivalent.

Similarly to scenario 4, scenario 5 also included the Multi-scale Five model. This model again outperforms every other model but since it's based on unfair assumptions it can't be included in the final conclusion. One particularity of note for the Multi-scale Five model is that the gap between the mean headway KPI of the Multi-scale Five and second best model, IDM, was reduced from 6.1% to 3.6% in scenario 5 compared to scenario 4. This reduction in gap is further explored in the next chapter 8 Discussion & Recommendation.

Model	KPI	t-value	df	p-value
Multi-scale Five	$SMAPE_x$	2.407	33.9	0.011
vs	$SMAPE_v$	1.123	52.6	0.133
IDM	$SMAPE_h$	3.337	41.5	<0.001
Multi-scale Five	$SMAPE_x$	2.960	40.5	0.003
vs	$SMAPE_v$	1.767	59.4	0.041
TDIDM	$SMAPE_h$	3.808	48.8	<0.001
Multi-scale Five	$SMAPE_x$	2.103	33.2	0.022
vs	$SMAPE_v$	1.436	51.4	0.078
Multi-scale	$SMAPE_h$	3.221	39.6	<0.001

Table 7.4: t-test results UNINA scenario 4 (no NDRT).

The differences in performance of the various models which were noted in the above section also need to be tested and statistically proven to be significant. This is done following the methodology described in chapter 6.1.4 Assessment Criteria. The above table gives an overview of the performed t-tests.

Since all the baseline models performed nearly equivalent in scenario 4 there is no need to perform a t-test for those models. The only t-test performed for scenario 4 are for the Multi-scale Five model. These results show that the Multi-scale Five model can be considered better than the other models for nearly every KPI. The only areas where the Multi-scale Five model seem to not be significantly better is in its speed estimations.
Another particularity of note is that, despite being the best performing model for headway in scenario 4, the Multi-scale Five vs TDIDM model t-test has the highest t-value. This is due to the low std of the TDIDM model compared to the other baseline models which reduced its probability of being statistically greater than the Multi-scale Five model.

Model	KPI	t-value	df	p-value
Multi-scale Five	$SMAPE_x$	2.434	57.2	0.009
vs	$SMAPE_v$	1.238	57.8	0.110
IDM	$SMAPE_h$	3.893	50.0	<0.001
Multi-scale Five	$SMAPE_x$	2.518	60.6	0.007
vs	$SMAPE_v$	0.752	60.1	0.227
TDIDM	$SMAPE_h$	3.905	44.5	<0.001
Multi-scale Five	$SMAPE_x$	1.757	57.1	0.042
vs	$SMAPE_v$	0.494	60.2	0.312
Multi-scale	$SMAPE_h$	3.977	43.5	<0.001
Multi-scale Five	$SMAPE_x$	1.954	43.3	0.029
vs	$SMAPE_v$	1.46	46.2	0.076
Distraction model	$SMAPE_h$	3.479	47.1	0.001
Distraction model	$SMAPE_x$	-0.246	51.4	0.597
vs	$SMAPE_v$	-0.491	54.3	0.687
IDM	$SMAPE_h$	0.106	61.3	0.458
Distraction model	$SMAPE_x$	-0.346	47.0	0.634
vs	$SMAPE_v$	-0.885	51.4	0.81
TDIDM	$SMAPE_h$	0.565	61.4	0.287
Distraction model	$SMAPE_x$	-0.684	51.5	0.751
vs	$SMAPE_v$	-1.065	51.2	0.854
Multi-scale	$SMAPE_h$	0.727	60.7	0.235

Table 7.5: t-test results UNINA scenario 5 (yes NDRT).

The last table in this section shows the t-test results of the various models used in scenario 5. The t-tests of the Multi-scale Five model in scenario 5 are similar to those in scenario 4. The only area where the Multi-scale model isn't significantly better than the other models is in its speed predictions. Other than that all other KPI measurements show that the Multi-scale Five model is significantly better than the other models, including the Distraction model. That said compared to the results in scenario 4 the t-values are a bit lower but still far above the limit of the 95% null hypothesis.
The second half of the table shows the t-tests which compare the Distraction model with the other models. Despite having a better mean headway KPI value and smaller confidence interval than the other models this difference is not found to be significant. Compared to the other models the Distraction model only has a ~75% chance of being better in headway prediction than the other HF baseline models. Moreover when performing the tests with the IDM the results are similar. In conclusion all four models, IDM, TDIDM, Multi-scale, and distraction are statistically equivalent for all KPIs.

7.3.2. CARRS-Q Data

The assessment of the individual trajectory calibration of the second dataset, CARRS-Q, is done in the same manner as the UNINA dataset. The tables below, 7.6, 7.7, 7.8, 7.9, summarise the results of the three KPIs used in the assessment and their statistical significance.

Model	Mean	Median	Min	Max	Std
IDM					
$SMAPE_x$ [m]	98.7%	99.2%	95.9%	99.7%	1.0%
$SMAPE_v$ [m/s]	96.6%	96.9%	91.9%	98.5%	1.4%
$SMAPE_h$ [s]	94.4%	95.3%	86.7%	97.7%	2.5%
TDIDM					
$SMAPE_x$ [m]	98.2%	98.4%	94.9%	99.9%	1.4%
$SMAPE_v$ [m/s]	95.5%	95.9%	90.8%	98.6%	2.3%
$SMAPE_h$ [s]	92.5%	93.5%	83.9%	97.9%	3.8%
Multi-scale					
$SMAPE_x$ [m]	98.4%	98.6%	96.0%	99.8%	0.9%
$SMAPE_v$ [m/s]	95.6%	96.1%	90.8%	98.4%	1.8%
$SMAPE_h$ [s]	93.2%	93.4%	88.7%	97.9%	2.2%
Multi-scale Five					
$SMAPE_x$ [m]	98.8%	99.0%	96.9%	99.7%	0.6%
$SMAPE_v$ [m/s]	96.3%	96.6%	92.5%	97.8%	1.2%
$SMAPE_h$ [s]	95.1%	95.5%	91.4%	96.9%	1.4%

Table 7.6: Individual trajectory calibration SMAPE results CARRS-Q scenario 1 (no NDRT).

The results of scenario 1, the baseline scenario, show that the best performing baseline model is the IDM model. It has a higher mean value for all KPIs over the other two models while having an equal or lower std. This result is a bit surprising since TDIDM model was specifically built for this dataset and performs the worst in this test. This isn't inline with the results from the original paper but the original paper doesn't cover the individual trajectory calibration results and only covers the calibrated model results which might be the cause for these discrepancies. Another particularity of note is that the Multi-scale model has the best negative outliers which also causes it to have slightly better std than the other models.

Lastly when the Multi-scale Five model is added to the comparison it comes out on top. It provides the best mean values with much lower std. This is especially the case for the headway where it not only has the best mean headway KPI but also nearly no outliers. That said when these results are compared to the data obtained using the UNINA data it is noticeable that the gap between the Multi-scale Five model and the baseline models has decreased. In general when comparing the UNINA data and the CARRS-Q data there is a noticeable increase in performance when using the CARRS-Q dataset but the increase for the baseline models is about twice as large as the increase for the Multi-scale Five model.

Model	Mean	Median	Min	Max	Std
IDM					
$SMAPE_x$ [m]	98.4%	98.8%	92.4%	99.4%	1.2%
$SMAPE_v$ [m/s]	95.8%	96.0%	88.9%	97.8%	1.7%
$SMAPE_h$ [s]	92.7%	93.9%	80.0%	96.1%	3.5%
TDIDM					
$SMAPE_x$ [m]	98.1%	98.3%	91.7%	99.7%	1.6%
$SMAPE_v$ [m/s]	94.8%	95.2%	86.4%	98.4%	2.4%
$SMAPE_h$ [s]	91.8%	92.9%	75.1%	96.8%	4.2%
Multi-scale					
$SMAPE_x$ [m]	98.4%	98.5%	96.5%	99.3%	0.6%
$SMAPE_v$ [m/s]	95.1%	95.3%	91.1%	97.0%	1.2%
$SMAPE_h$ [s]	92.7%	93.2%	83.2%	95.5%	2.6%
Distraction model					
$SMAPE_x$ [m]	98.3%	98.5%	95.9%	99.6%	0.7%
$SMAPE_v$ [m/s]	95.1%	95.3%	92.0%	96.9%	1.3%
$SMAPE_h$ [s]	94.1%	94.4%	87.0%	96.8%	2.2%
Multi-scale Five					
$SMAPE_x$ [m]	98.5%	98.6%	95.3%	99.7%	0.9%
$SMAPE_v$ [m/s]	95.4%	95.4%	91.6%	98.2%	1.3%
$SMAPE_h$ [s]	94.1%	94.6%	84.5%	97.7%	2.8%

Table 7.7: Individual trajectory calibration SMAPE results CARRS-Q scenario 3 (yes NDRT).

The following table, which covers the results of scenario 3, shows similar results to the results of scenario 1. The first unusual result is that the TDIDM model performs the worst of all the models, this is similar to what happened in scenario 1 but still not inline with the results of the original paper. That said when looking at the outliers it looks like TDIDM has one particularly bad outlier caused by the trajectory of participant 21. When this outlier is excluded, the performance results of TDIDM increase a bit and nearly match the performance of the other models. So when excluding the outlier all three baseline models perform the same and their only differentiating factor is their distribution or std. When looking purely at the standard distributions the best performing model is the Multi-scale model, its std is between 25% and 50% lower than those of the other models for the various KPI.
The results of the Multi-scale Five model are also included in the above table. Its shows that the Multi-scale Five model does perform a bit better than the other models and this improvement is a bit larger than the improvement found in scenario 1. In scenario one the improvement of the best baseline model, IDM, in the mean headway value is 0.7% while in scenario 3 this improvement is 1.4%. This change in improvement is mainly due to the loss of performance of the IDM model in scenario 3. This big loss, compared to the other models, is expected since the IDM model isn't made to support human factors so the distraction in scenario 3 causes more issues for IDM than for the other models which do support human factors.
Last are the results of the Distraction model. Overall the Distraction model is one of the best performing models, it is able to match the Multi-scale Five model in mean KPI and is even able to have a slightly lower std. Since the Distraction model is based on the Multi-scale model it is expected to be at least equal in the three KPIs. This is indeed the case, in terms of mean spacing and speed it is equal to the Multi-scale model. It even roughly matches the same outliers as the Multi-scale model, this is likely due to the fact that both models use the same base and have the same weaknesses. As for headway, that is the area where the Distraction model seems to perform best compared to the Multi-scale model. This is a bit odd since none of the changes to the Distraction model are specifically made with headway in mind, so having an improvement in only one KPI instead of all three is unexpected especially since headway depends on speed and spacing.

Model	KPI	t-value	df	p-value
Multi-scale Five	$SMAPE_x$	0.471	54.5	0.320
vs	$SMAPE_v$	-1.097	59.8	0.862
IDM	$SMAPE_h$	1.347	49.6	0.092
Multi-scale Five	$SMAPE_x$	2.512	44.3	0.008
vs	$SMAPE_v$	1.556	46.4	0.063
TDIDM	$SMAPE_h$	3.543	39.9	0.001
Multi-scale Five	$SMAPE_x$	2.289	56.2	0.013
vs	$SMAPE_v$	1.769	53.5	0.041
Multi-scale	$SMAPE_h$	4.038	53.4	<0.001
IDM	$SMAPE_x$	1.952	55.5	0.028
vs	$SMAPE_v$	2.238	52.1	0.015
TDIDM	$SMAPE_h$	2.299	53.9	0.013
IDM	$SMAPE_x$	1.521	61.8	0.067
vs	$SMAPE_v$	2.54	59.0	0.007
Multi-scale	$SMAPE_h$	2.028	61.1	0.023

Table 7.8: t-test results CARRS-Q scenario 1 (no NDRT).

With the individual results of the various models analysed, the models can now be statistically compared to one another to find the best performing model. The t-test results table of scenario one include t-tests between the Multi-scale Five model vs other models and of IDM vs other models. The first set is included since the Multi-scale Five model provides the theoretically best results and the second set since the IDM model performed surprisingly well compared to the other models.

From the t-test results of the Multi-scale Five model it is clear that the model is significantly better than the TDIDM and Multi-scale models especially when predicting the headway curves. The only area where the Multi-scale model has some issues is in its ability to predict the speed curves. This can be seen in its comparison against the TDIDM speed KPI where it's just under the limit of being significantly better, at 93.7%. That said the most unexpected result of this t-test is when the Multi-scale Five model is compared to the IDM model. It confirms what was already observed in the individual results above, that the IDM model is equivalent to the Multi-scale Five model in scenario 1. This is a major difference between the CARRS-Q data and the UNINA data. In the UNINA data the IDM was only equivalent on one of the KPI but in this dataset it is equivalent on all three KPIs. The reason for this is difficult to attribute to one cause and will be further elaborated in chapter 8 Discussion & Recommendation.

Since the IDM model performed so well for scenario 1, additional t-tests were performed to compare it to the other baseline models. These t-tests further confirm how well the IDM model performed as it is significantly better for nearly every KPI compared to the baseline models. In total the IDM is significantly better on five out of the six KPIs; this is the same result as the Multi-scale Five model but the one KPI which is not significantly better is different.

Model	KPI	t-value	df	p-value
Multi-scale Five	$SMAPE_x$	0.477	56.6	0.318
vs	$SMAPE_v$	-1.126	58.6	0.868
IDM	$SMAPE_h$	1.769	59.7	0.041
Multi-scale Five	$SMAPE_x$	1.493	49.3	0.071
vs	$SMAPE_v$	1.095	47.5	0.140
TDIDM	$SMAPE_h$	2.503	54.4	0.008
Multi-scale Five	$SMAPE_x$	0.971	53.2	0.168
vs	$SMAPE_v$	0.737	61.8	0.232
Multi-scale	$SMAPE_h$	1.999	61.6	0.025
Multi-scale Five	$SMAPE_x$	1.082	56.8	0.142
vs	$SMAPE_v$	-1.878	57.6	0.967
Distraction model	$SMAPE_h$	-0.097	58.9	0.539
Distraction model	$SMAPE_x$	-0.341	47.3	0.633
vs	$SMAPE_v$	0.544	46.4	0.294
IDM	$SMAPE_h$	2.005	53.1	0.025
Distraction model	$SMAPE_x$	0.879	41.6	0.192
vs	$SMAPE_v$	-0.116	62.0	0.546
TDIDM	$SMAPE_h$	2.738	47.3	0.004
Distraction model	$SMAPE_x$	-0.187	61.1	0.574
vs	$SMAPE_v$	0.847	61.9	0.200
Multi-scale	$SMAPE_h$	2.34	60.5	0.011

Table 7.9: t-test results CARRS-Q scenario 3 (yes NDRT).

The last table shows the statistical analysis results of the models in scenario 3. Since all the baseline models performed nearly the same in scenario 3 no additional t-tests were made specifically for the IDM model. Instead the Distraction model is compared to the baseline models.
The t-test results of the Multi-scale Five model for scenario 3 show quite a surprising result. According to this thesis's the assessment rules setup in 6.1.4 Assessment Criteria, this is the first time that the Multi-scale Five model can not be considered a better model than the baseline models. While it does perform significantly better on headway predictions than the other models this is the only KPI where the model performs better. This result is quite surprising since the Multi-scale Five model is supposed to be quite versatile since it doesn't follow the same assumptions and rule under which the baseline models operate. This could be caused by two things either the rule breaking logic of the Multi-scale Five model is unable to capture the effects present in scenario 3 or the baseline models found results much closer to the theoretical limit than the Multi-scale Five model did in the other scenarios.
In the second half of the table are the results of the Distraction model t-tests. Unlike with the UNINA data, the Distraction model is able to provide some significant improvements compared to the base models. These improvements are in the headway KPI as was already remarked in the individual results section. Surprisingly the certainty of the headway improvements are greater for the Distraction model than for the Multi-scale Five model. As for the speed and spacing KPIs the Distraction model is unable to provide any significant improvements over the IDM or TDIDM models. So according to the assessment rules this means that the model isn't better than the baseline models since only one out of three KPI is significantly greater.

7.4. Calibrated Model Results

This section covers the results and analysis of the calibrated models, it is divided into two parts. The first part is the calibrated model parameters calculated from the calibration data split and the second part is the calibrated model results calculated from the validation data split.

7.4.1. Calibrated Model Parameters

The tables 7.10, 7.11, 7.12, 7.13 below list the parameters used for each model and each seed. These parameters are calculated by taking the mean value of the parameters from the calibration data split.

This subsection won't fully analyse the parameters listed in the tables, instead the main focus of the analysis is comparing the yes and no NDRT datasets. This means comparing table 7.10 to 7.11 and table 7.12 to 7.13.

Parameter & Unit		Models & 3 Seeds								
		IDM			TDIDM			Multi-scale		
		1	2	3	1	2	3	1	2	3
b_{max}*	$[m/s^2]$	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0
a_{max}*	$[m/s^2]$	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0
b_{comf}	$[m/s^2]$	1.89	1.43	1.72	1.38	1.52	1.17	1.56	1.79	1.26
v_0	$[m/s]$	38.35	38.79	40.47	38.36	38.32	39.43	36.76	36.75	38.57
s_0	$[m]$	28.89	27.27	27.46	29.77	23.83	28.25	17.45	23.61	19.29
T	$[s]$	2.49	2.46	2.51	3.0	2.79	3.07	3.5	3.24	3.17
τ	$[s]$				0.23	0.25	0.18	0.45	0.56	0.31
δ	$[-]$				-1.03	-0.74	-0.98			
γ	$[-]$				2.28	2.65	2.16			
φ	$[-]$				0.28	0.38	0.31			
δ_{sign}	$[-]$							0.34	0.42	0.28
TD_d	$[-]$							0	0	0

Table 7.10: Average model parameters UNINA scenario 4 (no NDRT). *Fixed values, based on test data.

Parameter & Unit		Models & 3 Seeds											
		IDM			TDIDM			Multi-scale			Distract model		
		1	2	3	1	2	3	1	2	3	1	2	3
b_{max}*	$[m/s^2]$	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0
a_{max}*	$[m/s^2]$	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0
b_{comf}	$[m/s^2]$	4.27	4.39	4.09	2.79	3.01	2.07	4.11	3.91	3.71	4.07	3.94	3.64
v_0	$[m/s]$	37.92	38.83	39.56	35.83	37.03	37.34	36.74	37.55	38.07	35.05	35.84	36.24
s_0	$[m]$	10.09	8.19	8.48	25.25	24.93	27.18	9.39	8.07	7.86	9.89	8.26	9.87
T	$[s]$	3.59	3.77	3.61	3.59	3.83	3.69	4.03	4.0	4.1	4.36	4.38	4.14
τ	$[s]$				0.1	0.1	0.1	0.22	0.19	0.25	0.25	0.24	0.22
δ	$[-]$				-0.64	-0.85	-0.45						
γ	$[-]$				0.79	1.04	0.99						
φ	$[-]$				0.14	0.13	0.14						
δ_{sign}	$[-]$							0.21	0.27	0.25	0.15	0.2	0.15
TD_d	$[-]$							0.25	0.19	0.18	0.52	0.58	0.53
β_a	$[-]$										0.38	0.41	0.57

Table 7.11: Average model parameters UNINA scenario 5 (yes NDRT). *Fixed values, based on test data.

Overall when comparing the parameters of the two UNINA scenarios the differences seem to be inline with expectations. The differences between the two tables are mainly in the spacing at standstill (s_0), desired headway (T) and reaction time (τ).
The desired speed is the same between the two scenarios which makes sense since both scenarios are setup in the exact same manner. The only change in speed is seen in the Multi-scale based models. They seem to have slightly lower speeds than the IDM or TDIDM model but only by a couple of meters per seconds. For the desired spacing at standstill it seems to decrease in scenario 5, this is true for all models except for the TDIDM model. The cause of this isn't immediately obvious but it is likely that the drivers react a bit slower in the distracted scenario and consequently end up having a smaller minimum spacing. This reduction in minimum spacing is then negated by the next parameter, the desired headway. In the distracted scenario the desired headway increases by roughly one second which is inline with the task difficulty homeostasis theory from Fuller et al. (2007). Another particularity of note about the desired headway is that the TDIDM and Multi-scale model seem to estimate larger

headways than the IDM model, this is likely due to how human factors are integrated into the models resulting in larger headways. The last shared parameter is the reaction time. Overall the reaction time seems to be fairly consistent between the two scenarios but there are some differences between the two models. Most notably the reaction time in the Multi-scale model scenario 5 seems to be a bit lower than all other reaction times. Please note that when calculating the reaction time on the TDIDM model both the normal reaction time (τ) and additional reaction time (φ) need to be added up. The last parameter of interest is the distraction task difficulty (TD_d) from the Multi-scale and Distraction model in scenario 5. Since this parameter is only present in these two models in scenario 5 it can only be compared to one another and not to scenario 4. Overall the task difficulty seems to be much higher in the Distraction model, this does make sense since it is only applied at limited moments whereas the distraction in the Multi-scale model is continuously applied.

In conclusion the changes in parameters between scenario 4 and 5 are as expected and explainable. Furthermore there doesn't seem to be a large variability between the different seeds so it can be concluded that the different driver trajectories are fairly consistent in their behaviour and that the seeds weren't affected by outliers. Naturally there are a few exceptions to this, like the spacing at standstill for the Multi-scale model in scenario 4 seed 2 or the comfortable breaking acceleration for the TDIDM model in scenario 5 seed 3, but these are few and far between.

Parameter & Unit		Models & 3 Seeds								
		IDM			TDIDM			Multi-scale		
		1	2	3	1	2	3	1	2	3
b_{max}*	$[m/s^2]$	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0
a_{max}*	$[m/s^2]$	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0
b_{comf}	$[m/s^2]$	0.82	1.02	1.01	1.15	1.35	1.55	1.09	1.45	0.95
v_0	$[m/s]$	16.84	15.87	15.74	21.78	22.32	17.42	26.62	23.09	30.58
s_0	$[m]$	2.47	2.7	2.16	6.37	5.61	5.61	4.52	4.7	3.55
T	$[s]$	1.09	1.44	1.43	2.44	2.74	2.53	1.45	1.77	1.83
τ	$[s]$				0.18	0.21	0.19	0.37	0.41	0.35
δ	$[-]$				-1.78	-1.65	-1.66			
γ	$[-]$				1.75	1.46	1.9			
φ	$[-]$				0.3	0.29	0.32			
δ_{sign}	$[-]$							0.18	0.2	0.15
TD_d	$[-]$							0	0	0

Table 7.12: Average model parameters CARRS-Q scenario 1 (no NDRT). *Fixed values, based on test data.

Parameter & Unit		Models & 3 Seeds											
		IDM			TDIDM			Multi-scale			Distract model		
		1	2	3	1	2	3	1	2	3	1	2	3
b_{max}*	$[m/s^2]$	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0	8.0
a_{max}*	$[m/s^2]$	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0
b_{comf}	$[m/s^2]$	1.41	2.23	2.1	1.23	1.41	1.5	0.93	0.93	0.92	1.27	1.67	1.42
v_0	$[m/s]$	12.7	13.95	14.15	24.53	23.35	22.4	35.07	33.89	33.62	33.12	30.5	30.82
s_0	$[m]$	4.36	4.97	3.14	7.4	6.63	6.52	4.65	4.45	4.18	5.31	5.48	4.79
T	$[s]$	1.34	1.64	1.55	2.02	2.46	2.53	2.17	2.62	2.29	2.14	2.56	2.25
τ	$[s]$				0.14	0.12	0.14	0.19	0.21	0.21	0.31	0.36	0.36
δ	$[-]$				-1.0	-0.79	-1.24						
γ	$[-]$				1.31	1.18	1.45						
φ	$[-]$				0.33	0.28	0.32						
δ_{sign}	$[-]$							0.2	0.18	0.18	0.18	0.18	0.22
TD_d	$[-]$							0.34	0.37	0.3	0.53	0.5	0.53
TD_{dsp}	$[-]$										0.3	0.41	0.26
β_a	$[-]$										0.5	0.42	0.61

Table 7.13: Average model parameters CARRS-Q scenario 3 (yes NDRT). *Fixed values, based on test data.

The second set of tables, 7.12 and 7.13 show the calibrated model parameters for the CARRS-Q dataset. Overall it seems that there is more variation and contrast, in the parameters, between the various models and seeds for the CARRS-Q than in the UNINA dataset. The exact differences between the two scenarios and two dataset are explained below.
To start the desired speed is a lot less stable. Not only are there differences in desired speed between the two scenarios, there are also large differences in desired speed between the various models. For example in scenario 3 the IDM model predicts a desired speed of roughly 13 m/s, which seems to match the average speed in the scenario, but the Multi-scale model estimates a desired speed of 33 m/s. Such difference also exist in scenario 1 but they are far less extreme. While there are possible explanations for this phenomenon, which have to do with the task demand in the Multi-scale model, they are difficult to confirm and will be discussed in chapter 8 Discussion & Recommendation. On the other hand a parameter with much less variability is the desired spacing at standstill. While there are some differences between the various model ranging from 4 to 7 meters these do stay consistent between scenario 1 and 3. The model with the highest spacing is the TDIDM model which was also the case for the UNINA dataset and the IDM having the smallest spacing is also consistent between the two datasets. Another parameter whose changes are consistent with the UNINA dataset is the headway. Overall the HF models seem to have a desired headway which is higher than the IDM model; that increase can be explained for their desire to accommodate for the reduction in safety caused by HFs. Furthermore when the drivers driving difficulty is increased by the distraction in scenario 3 another increase in headway is seen across all models. The last shared parameter is the reaction time. Overall the reaction time seems to be a bit lower in the CARRS-Q dataset compared to the UNINA dataset, from roughly 0.5 to 0.4 seconds. Other than that it seems consistent, the reaction time for the Multi-scale models is a bit lower than that of the TDIDM model and the reaction time doesn't change between scenarios. The only anomaly seems to be the reaction time of the Multi-scale model in scenario 3 which is around 0.2 seconds. This might be due to the model trying to rapidly adapt for the sudden change in speed caused by the phone call but none of the other models have shown such behaviour so the exact cause is unknown. The last parameter which is discussed is only shared by the Multi-scale models, namely the distraction task demand. Overall the continuous distraction level of the Distraction model seems to be higher than the distraction level of the Multi-scale model. That said one peculiar anomaly is the spontaneous distraction level of the Distraction model. It would be expected that the drivers experience a higher distraction level when the phone first starts ringing but it would seem that the distraction level is lower on average. Looking further into the individual results it would seem that some people have little to no distraction task demand due the ringing phone while other have a strong reaction. This ends up averaging at a task demand just below the continuous task demand of the phone conversation. The reason for this division in task demand is further explored in

the discussion.

One last major difference between the two datasets is the variability seen when using different data splits. While the UNINA data was fairly consistent between the different seeds the CARRS-Q data seem to be a bit less stable. In the CARRS-Q dataset the different data splits seem to cause the parameters to show a variation in the order of 20 to 30% between the minimum and maximum. This variation in the UNINA dataset is roughly 5 to 10%. This would indicate that the drivers in the CARRS-Q dataset are more varied and less consistent with each other. This seems a bit unusual at first glance since the CARRS-Q dataset is much more homogeneous with only young drivers but this might be the exact cause of the problem. Young drivers usually have less experience and less experienced drivers show more variance in their driving behaviour.

7.4.2. Calibrated Model Performance

The follow subsection covers the results of the calibrated models. This includes both the average KPI results, the statistical t-tests and a more in-depth look at some of the individual results.

Scenario & KPI	Models & 3 Seeds											
	IDM			TDIDM			Multi-scale			Distract model		
	1	2	3	1	2	3	1	2	3	1	2	3
UNINA 4												
$SMAPE_x$ [m]	96.6%	97.0%	91.3%	86.8%	93.3%	88.0%	90.2%	96.9%	91.3%			
$SMAPE_v$ [m/s]	95.1%	96.0%	93.5%	94.1%	95.2%	92.9%	94.6%	95.7%	93.3%			
$SMAPE_h$ [s]	75.4%	78.6%	70.6%	61.8%	71.2%	62.0%	71.2%	79.2%	71.2%			
UNINA 5												
$SMAPE_x$ [m]	98.0%	98.2%	98.0%	97.9%	98.2%	97.9%	98.1%	98.3%	98.0%	97.6%	98.0%	98.0%
$SMAPE_v$ [m/s]	91.5%	91.4%	91.5%	91.5%	91.4%	91.5%	91.7%	91.5%	91.3%	90.1%	90.4%	90.8%
$SMAPE_h$ [s]	78.9%	80.5%	79.3%	78.8%	79.0%	78.6%	78.8%	80.5%	78.9%	77.3%	79.6%	78.8%
CARRS-Q 1												
$SMAPE_x$ [m]	95.5%	97.1%	96.7%	94.4%	95.9%	95.3%	96.2%	97.3%	97.1%			
$SMAPE_v$ [m/s]	94.8%	96.0%	95.6%	90.1%	92.9%	89.2%	94.5%	95.3%	95.2%			
$SMAPE_h$ [s]	75.7%	81.8%	83.6%	73.6%	78.2%	78.0%	81.2%	82.6%	84.7%			
CARRS-Q 3												
$SMAPE_x$ [m]	95.9%	96.9%	96.1%	93.5%	95.8%	94.4%	95.8%	96.9%	96.9%	95.7%	96.8%	96.5%
$SMAPE_v$ [m/s]	94.9%	95.0%	94.7%	90.2%	92.6%	90.8%	93.6%	94.6%	94.4%	93.1%	93.8%	93.7%
$SMAPE_h$ [s]	80.1%	81.7%	80.7%	69.9%	79.6%	73.7%	78.8%	80.9%	84.0%	76.6%	80.5%	79.6%

Table 7.14: Calibrated model average KPI, all scenario's all seeds.

Scenario & KPI	Distraction model vs								
	IDM			TDIDM			Multi-scale		
	1	2	3	1	2	3	1	2	3
UNINA 5									
$SMAPE_x$ [m]	0.871	0.734	0.550	0.811	0.737	0.448	0.925	0.837	0.610
$SMAPE_v$ [m/s]	0.873	0.832	0.761	0.877	0.813	0.773	0.912	0.847	0.688
$SMAPE_h$ [s]	0.754	0.688	0.575	0.741	0.402	0.469	0.748	0.684	0.511
CARRS-Q 3									
$SMAPE_x$ [m]	0.590	0.552	0.274	**0.016***	**0.033***	**0.002***	0.577	0.582	0.761
$SMAPE_v$ [m/s]	0.951	0.999	0.922	**0.004***	**0.005***	**0.001***	0.672	0.973	0.846
$SMAPE_h$ [s]	0.833	0.620	0.612	**0.027***	0.397	**0.049***	0.760	0.539	0.863

Table 7.15: Calibrated model t-test, Distraction model vs other models - all seeds. * Passed the 5% probability threshold.

Table 7.14 shows the average KPIs of the validation data split for the calibrated models. This is one

combined overview which shows all models, all scenarios and all seeds. An example of an average performing calibrated trajectory for both UNINA and CARRS-Q can be seen in figures 7.5 and 7.6. When compared with the KPIs of the individual trajectory results, the average values of the calibrated models are much lower. This is to be expected since generalised calibrated models score worse on known data but better on unknown data than individual trajectory models. On average the speed and spacing KPIs of all models lost about 5% compared to individual trajectory KPIs. This loss in performance is even greater for the headway KPI, most models lost around 10% and the Distraction model lost about 15% in all scenarios. This additional loss in performance by the Distraction model doesn't seem to be arbitrary since the headway was also the best performing KPI of the Distraction model compared to the baseline models. It is likely that the headway estimation of the Distraction model was overfit for the individual trajectories and this over-fitting was lost when the model parameters were averaged.
In the end it would seem that all models are roughly equal. For some scenarios some of the models score 1 or 2% better than the Distraction model but this depends on the seed so it is difficult to draw a strong conclusion. Most models aren't too sensitive to the changes in seed, and subsequently different data splits, but the exception to this seems to be the headway of the TDIDM model. Where most models have a KPI difference of at most 5% due to changes in seed for the TDIDM headway this difference is between 5 and 10% for every scenario.

Using the KPI results of the validation data split a final set of t-tests are performed. These t-tests compare the Distraction model against the other models for both scenarios with distraction for all seeds. The results of these t-tests can be seen in table 7.15.
The results show that all the models are statistically equivalent for the UNINA dataset. This is the same conclusion as was found for the individual trajectories. So the Distraction model is not better at predicting driver behaviour than the baseline models when the driver is experiencing continuous mental/visual distractions.
The CARRS-Q dataset does offer a slightly more diverse set of results. The t-tests show that the Distraction model is not always equivalent to the baseline models. To start the Distraction model is almost always significantly better than the TDIDM with the exception of the headway KPI in seed 2. According to the assessment rules setup in chapter 6.1.4 Assessment Criteria if two out of three KPIs, of the Distraction model, are significantly better and the third KPI is equivalent it means that the Distraction model is significantly better than the other model. This means that the Distraction model is significantly better than the TDIDM model for all three seeds. When comparing the Distraction model to the IDM and Multi-scale models the t-tests show that the results of the spacing and headway KPI are equivalent. As for the speed KPI it shows that in most cases the Distraction model is significantly worse. Since only one of the KPI is significantly worse and the other two are equivalent it can't be said that the Distraction model is significantly worse than the IDM or Multi-scale model. Overall these results differ quite a bit from the individual trajectory results. In the individual trajectory results the Distraction model performed significantly better on headway than the other models but this performance was lost, instead it gained a significantly better performance on speed and spacing over the TDIDM model.

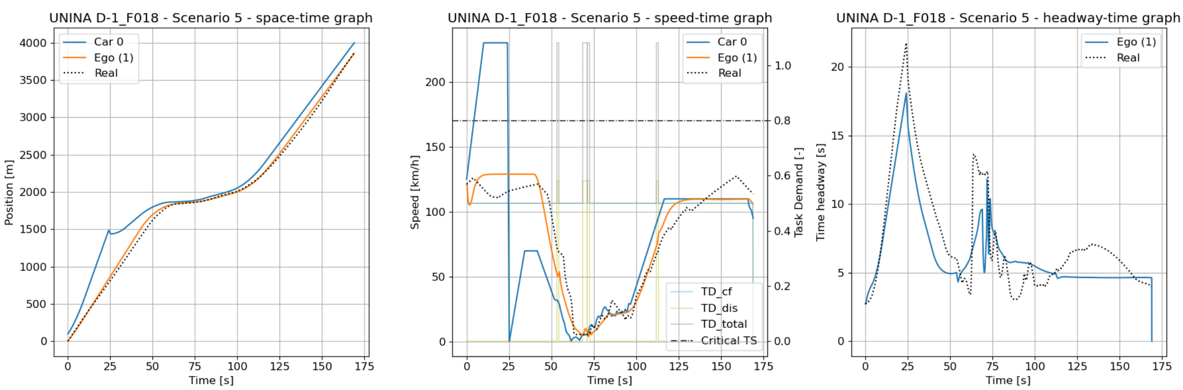

Figure 7.5: UNINA Seed 2, calibrated Distraction model single trajectory.

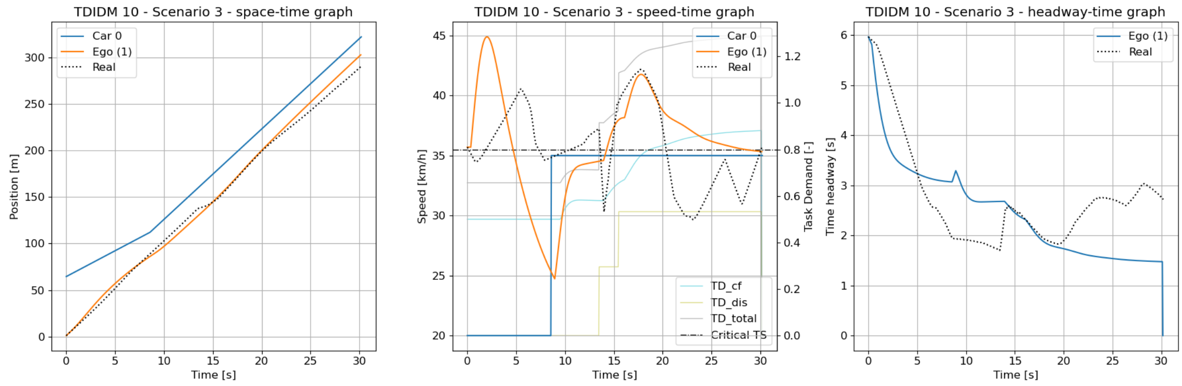

Figure 7.6: CARRS-Q Seed 1, calibrated Distraction model single trajectory.

7.4.3. Simulated Distraction Effect Analysis

The global model performance analysis given in the previous section gives us a good idea of which model performed best but it doesn't explain why this is the case. To this end a more detailed analysis is needed. To do this some of the individual validation trajectories of the calibrated models are analysed in more detail.

The trajectories which were analysed are from data-split distracted scenarios seed 1. Representative trajectories of each dataset are presented below.

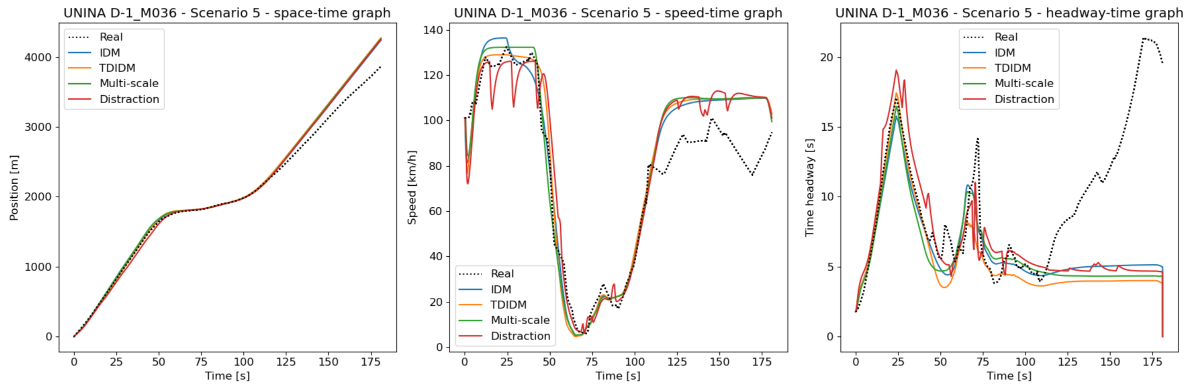

Figure 7.7: Trajectories of all models, UNINA validation example.

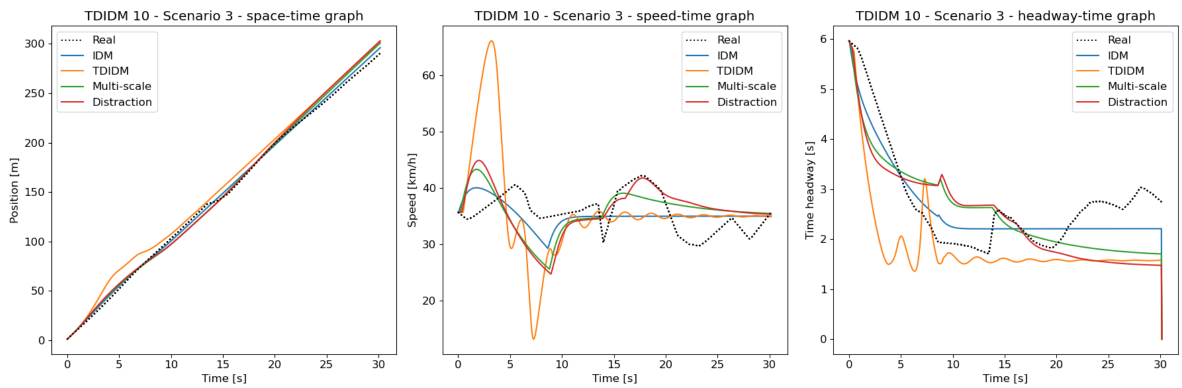

Figure 7.8: Trajectories of all models, CARRS-Q validation example 1.

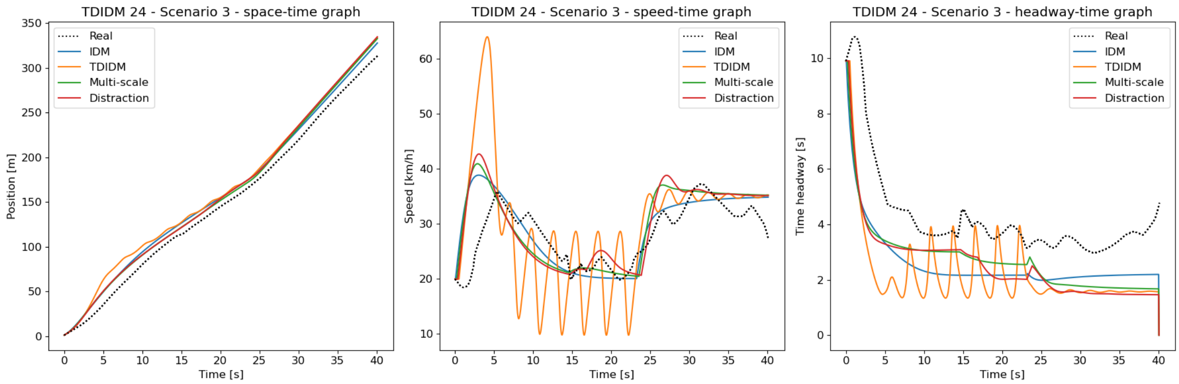

Figure 7.9: Trajectories of all models, CARRS-Q validation example 2.

The first model which is analysed is the IDM model. On average this model performed the best in terms of KPI and an explanation for this behaviour can be seen in the above graphs. The IDM is overall the most stable model, this is both a positive and a negative aspect. This stability means that there are few fluctuations, big or small, in the driving behaviour of the model. This results in really straight and consistent speed and headway curves. A good example of this can be seen in figure 7.8 and 7.9. These stable curves result in good scores for the KPI calculations since its never too far of the real trajectory. That said stable curves are contrary to the objective of HF models. The goal of HF factor models is to explain the dynamic fluctuations in driving behaviour seen in real trajectories. Naturally this is difficulty to achieve with a calibrated model since a calibrated model model can only capture general trends and not individual specific trends.
Another positive aspect of the IDM model is that it has the smallest overshoots. When strong accelerations or decelerations periods occur all models tend to overshoot the performance of the real trajectory, IDM model included. In reality the driver seems to anticipate the end of the acceleration periods and reacts pre-emptively or near instantaneously. Since the models aren't built with this type of anticipation in mind it results in overshoots. That said the IDM model reacts and recovers the fastest. This is likely due to the lack of reaction time in the IDM model leading to instantaneous changes in accelerations at the end of the acceleration periods. An example of this can be seen in the speed graph of figure 7.9 at 25 seconds.

The model which performed the worst in terms of KPI is the TDIDM model. The reason for this is quite clear when analysing the individual trajectories. Simply put the TDIDM model is too sensitive and unstable. The performance of the TDIDM model can be separated into three cases. In the first case the model performs well without any instability, see figure 7.7. Its performance is similar to the IDM model and in the example it is even slightly better. That said these stable cases are quite rare; in the second and majority of the cases the TDIDM model reacts as seen in figure 7.8. In these cases a large spike in speed at the start of the simulation causes the model to be offset for the rest of the simulation. The model tends to recover relatively slowly which causes a large impact on the KPI result of the model. In the third most extreme and least common case the model is completely unstable, see figure 7.9. The model simply never recovers from its initial mistakes and sometimes even amplifies the issue resulting in a terrible KPI. The reason why this model is so unstable is not entirely clear. It doesn't seem to be related to the calibration process as multiple methods of calibration were tested yielding no improvements; the issue seems to be linked to the inherent nature of the model.
In terms of HF the model is slightly better than the IDM model. The model is able to react to some of the smaller fluctuations in driving behaviour but it doesn't react to every fluctuation and when it does it greatly overestimates the response, see figure 7.8 speed curve around 6 seconds as example.

Compared to the previous two models the Multi-sale model behaves much closer to what would be expected of a HF model. In both the UNINA and CARRS-Q datasets it is able to predict some of the smaller fluctuations in driving behaviour. In the UNINA trajectories this is mainly visible in the headway curves. It seems to follow the trends in trajectory better than the IDM model though the differences aren't large. The CARRS-Q trajectories is where the Multi-scale model shows a major difference compared to the IDM model. Since the Multi-scale model knows when the phone calls start it is able to change its behaviour at that moment and this change in behaviour results in a notable jump in the speed and headway curves. These reactions are quite similar to the reactions seen in the real trajectory and are not captured at all in the IDM model. While these reactions happen at the correct moment and are of the correct type they have the wrong amplitude. As a result the trajectories of the real car and simulated car become slightly offset resulting in worse KPI for the rest of the simulation. This leads to the second big difference when compared to the performance of the IDM model. The general performance of the Multi-scale model is a bit worse than the IDM model; even without taking the offsets into account it seems that the Multi-scale model is a bit more conservative in its headway estimates. This can be beneficial as seen in 7.9 or detrimental as seen in 7.8 and on average this seems to be detrimental as reflected by the performance KPI. So in conclusion, the Multi-scale model is better at estimating the small fluctuations in driving performance than the previous two models but at the cost of general accuracy.

The last model is the Distraction model. With its more specific distraction triggers and distraction effects it is expected that the model is able to accurately capture most of the small fluctuations in driving behaviour while also keeping a good performance of the overall picture.
Looking at the trajectories of the UNINA dataset this expectation seems to be met. An example of this can be seen in the speed graph of figure 7.7. It shows that the model is able to estimate the spikes in speed in the first third of the trajectory, even if their amplitudes are on the low side. A similar performance can be seen in the last third of the trajectory where it is able to predict changes in behaviour at the right moment but the amplitude and type of reaction are incorrectly estimated. The middle third of the trajectory shows that the Distraction model has more difficulty estimating the low speed distraction effect. Instead of resulting in a decrease in speed it instead results in an increase in speed. These types of predicted effects shown in the example happen in most of the trajectories of the validation set. Similar behaviour is also seen in the CARRS-Q dataset. The Distraction model behaves quite similarly to the Multi-scale model but its two different distraction levels allows it to more accurately estimate the changes in speed that occur at and after the phone call. This can be seen in figure 7.8 around the second 20 mark and in figure 7.9 at the 18 second mark. That said this increase in performance again comes at the price of general performance.

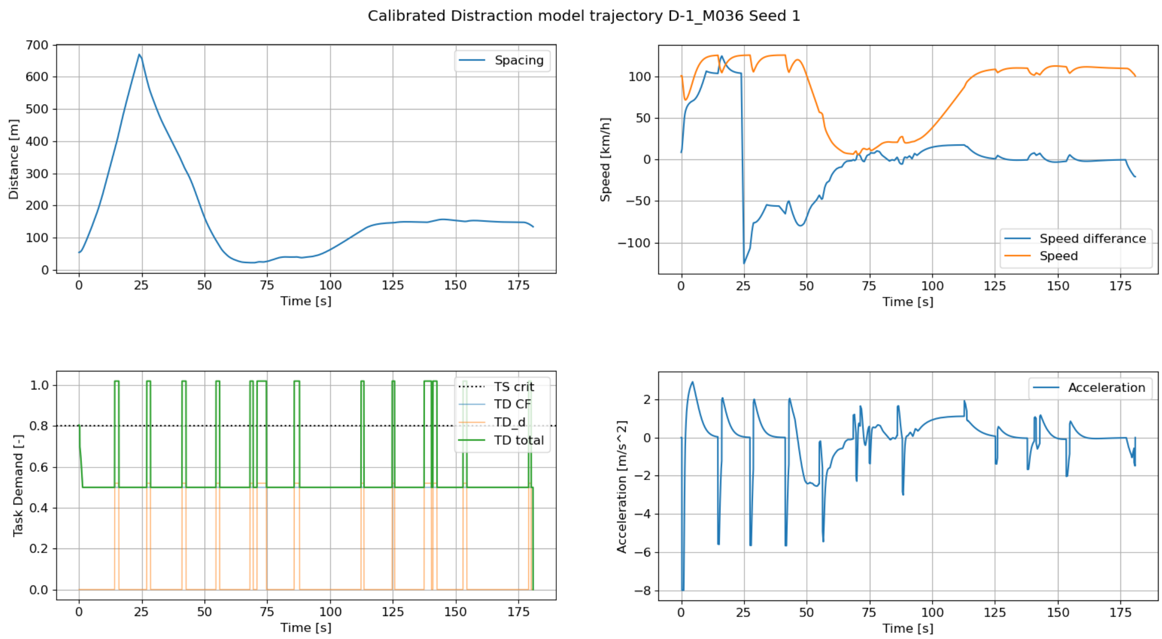

Figure 7.10: Detailed calibrated Distraction model trajectory - UNINA M036 Seed 1.

To better understand why the calibrated distraction model behaves in this way, a trajectory is further analysed. The UNINA trajectory in figure 7.7 is used as example for this purpose, more details regarding this trajectory are given in figure 7.10. The parameters of the calibrated Distraction model are given in table 7.16.

Parameter	Value	Unit
b_{max}	8	m/s^2
a_{max}	4	m/s^2
b_{comf}	4.07	m/s^2
v_0	35.05	m/s
s_0	9.89	m
T	4.36	s
τ	0.25	s
δ_{sign}	0	$-$
TD_d	0.52	$-$
β_a	0.38	$-$

Table 7.16: Parameters for calibrated Distraction model Seed 1.

From these figures a few key observations can be made. The task demand for car following stays constant at the minimum value of 0.5 during the whole simulation. This is one of the downsides of the UNINA dataset, due to its high headways of roughly 4 seconds it never breaches the h_0 threshold of 3 seconds so the TD_{cf} never increases beyond the minimum. The other part of task demand comes from the distraction, this is a constant value of 0.52, so when the two are summed they have a task demand of 1.02 which exceeds the TS_{crit} threshold. This means that under normal circumstances no perception modifiers are applied but when distracted a 20% reduction is applied to desired speed and desired headway, a 10% increase in reaction time, perceived speed difference and perceived spacing, and a 20% reduction in maximum acceleration.

The acceleration graph in figure 7.10 has multiple spikes in acceleration. Most of these spikes start with a strong deceleration and have a subsequent acceleration. When comparing this graph with the task demand graph it becomes obvious that these spikes are caused by the distractions. Additionally when carefully overlapping these two graphs it is clear that the duration of the acceleration is also tied to the duration of the distraction. The deceleration starts about half a second after the start of the distraction, this matches the modified reaction time $(0.25s + 10\% * 2s)$ and the acceleration starts half a second after the end of the distraction. That said not all spikes seem to have the same behaviour. In total three different kinds of spikes can be identified. The first are large-amplitude deceleration spikes, the second are short-amplitude acceleration spikes and the last are short-amplitude deceleration spikes. These can be observed in the same order in the acceleration graph. This means that the same distraction can cause multiple types of changes in driving behaviour. The deciding factor for the exact reaction seems to be tied to two variables, the speed of the vehicle and the spacing/headway to the leader.
The large-amplitude deceleration spikes happen at high speeds, close to desired speed, and high spacing, see the first third of the trajectory. The short-amplitude acceleration spikes happen at low speeds and low spacing while the short-amplitude deceleration spikes happen at high speeds and low spacing, see the second and last third of the trajectory. The first two behaviours can be explained with their corresponding circumstances but the last one is more subtle. The easiest to explain is the short-amplitude acceleration spike, keep in mind that the distraction reduces the desired speed and desired headway. Since the car is driving at speeds far below the desired speed this desired speed reduction has nearly no effect, only the decrease in desired headway has an effect. The end result is an acceleration to reach the new shorter desired headway and naturally once the distraction ends and the desired headway is back to normal the car decelerates. A similar explanation can be given for the large-amplitude deceleration spikes but in this case it's the reduction in desired speed which plays a dominant role. The reduction in desired speed causes the vehicle to be above the new desired speed so the vehicle decelerates. Once the distraction is over they accelerate to reach the original desired speed. The small-amplitude deceleration spikes can be explained with the same logic but there is a subtle difference. During the large-amplitude spikes the vehicle drives at ∼115 km/h while the reduced desired speed is ∼100 km/h but during the small-amplitude spikes the vehicle drives at ∼105 km/h.

This means that the vehicle needs to reduce its speed by 15 km/h in the first scenario and 5 km/h in the second scenario hence the difference in amplitude of acceleration.

8

Discussion & Recommendation

The discussion covers all alternatives considered in the methodology and all found anomalies in the data or results. For each point the possible impact on the results will be explained and if such impact is unknown or unclear a recommendation for how this impact could be quantified is made.
The discussion is split into section which follow the same chronological order as the thesis.

8.1. Methodology

The middle chapters of this thesis make quite a few trade-offs and decisions in order to decide on the methodology. Most of these decisions are based on literature or small tests and only rarely are decisions based on intuition. This section provides some insight into these decisions and provides additional information about the possible alternatives.

The first important decision that is explored comes from the distraction framework, namely the distraction lifecycle structure. The lifecycle is divided into three stages based on the analysis and interpretation of the distraction literature. These three stages each play their own role with a bit of overlap at their edges. That said one of the question that arises from this chapter is: why three stages? It should be possible to achieve a similar result with two or even four stages. To start, the reason why a framework with distinct stages was chosen was to provide concrete steps that go from a global distraction to unique individual driver responses. A possible two-stage approach could be to have a distraction trigger and a merged intensity and effects. This approach would be easier to implement and understand but it reduces the flexibility of the framework. The opposite is true with a four-stage approach. It would increase flexibility but the increase in overlap between the stages would also reduce its clarity. This is why an approach with three stages was chosen; it provides a good balance between flexibility and clarity. To give an example of this, the reason why intensity and effect were split is to separate the driver response into two parts. First is the response in terms of meta parameters and second the response in terms of driver behaviour. This allows for more flexibility by having different meta-response levels trigger different effects. This comes at the cost of clarity since the difference and relation between meta parameter and driving behaviour is not always clear. If future research uncovers a new distraction mechanisms which can't be modelled by these existing stages it would be possible to add a new stage to the distraction lifecycle but based on the current literature the provided stages are the best solution.

The second trade-off which was made was for the base car-following model. Since the distraction framework is not a car-following model it needs to be integrated into a CF model. To this end an MCA was made to choose an appropriate existing CF model. An alternative would have been to design a new custom car-following model from scratch. Making a new CF model would allow for some more flexibility but it would take a lot of time and testing. One use of this flexibility would be to have a more complete integration of gaze data in the CF model, this could be done on a more tactical level similar to the research by Radhakrishnan et al. (2023) or on an operational level which would be new in this field. Perhaps by utilising a CF model that fully integrates gaze data, instead of only including gaze data in the distractions, it would lead to better distraction modelling performance and more realistic driving behaviour. Naturally this is only speculation and building and testing such a model would take a lot of time and likely only lead to marginal improvements. This hypothesis is supported by a CF model review performed by Saifuzzaman and Zheng (2014) which finds that most of the older models are quite robust and while improvement is possible most of the newer models only provide marginal improvements or improvements under specific conditions. Based on this it was deemed that building a fully new CF model was outside the scope of this thesis and that it would be better to use an existing proven CF model as a base. This is also the reason why modifications to the model were kept to a minimum, outside of the distraction portions, to avoid having to re-test and validate the CF parts of the model.

Further in the methodology the arguably most important decision is made. This is the approach for the validation methodology. The main research question is written with all types of distractions in mind. Validating such a question would naturally require large amounts of data to cover as many distraction types as possible. This wasn't possible for this thesis and the scope was instead limited to two datasets. The reason for this decision and limited scope validation approach can be found in 6.1.1 Proof of Concept for Validation. What isn't discussed in that chapter is the alternative/original validation approach which should be taken if enough data were available. For the most part the validation is the same but the focus would shift from the performance of the various models to how well each distraction can be represented by the distraction framework implementation. An additional step would be added to test the robustness of the distraction framework. This is done by implementing each distraction and measuring how well they perform, both in terms of correctness of the conceptual/theoretical implementation and in terms of distraction modelling performance. These tests for multiple distractions would naturally mean that a library of lifecycle stage building blocks is created which can then be used to build new kinds of distractions in future studies.

Another important decision-making point in the methodology is the final form of the Distraction model. The distraction framework offers various options to implement different types of distraction. In order to decide on the exact formulation of the three lifecycle stages multiple variants were made and tried. These variants were inspired by literature and trial and error, see 6.2.1 Distraction Model Variants. Naturally not all possible variants were tried only the most promising and relevant variants were tried and only the best result was kept in the final formulation of the implementation. The "most relevant" part is where this decision has ties with the previous discussion point. Since the scope of distraction types was limited quite a few variants were discarded because they weren't relevant for the limited scope. This doesn't mean that these variants were lacking in performance or unpromising but simply not appropriate for the chosen distractions. This leads to an interesting point of research, namely how many variants or building blocks are distraction specific and which ones can be used for multiple types of distractions. Since the distraction framework is built around the idea that different distractions with the same low-level characteristics can use the same building blocks to represent them it is important to further validate this idea. This wasn't done in this thesis due to the aforementioned lack of data so instead it is recommended that this is studied in future research.

One last trade-off which is covered in the methodology discussion is the choice of calibration algorithm. This thesis uses a genetic algorithm to calibrate the various models for the various dataset scenarios. Genetic algorithms are a type of stochastic search and optimization algorithm. These algorithms rely on randomness to explore the different solutions and optimize along the most promising found paths. This makes these types of algorithms capable and efficient at solving complex optimization problems. While they can't find the exact global minima they can consistently get close depending on the level of complexity (Spall, 2005). Genetic algorithms are one of the most well known algorithms of this type but other popular examples are the simulated annealing, particle swarm optimization. Simulated annealing functions by moving a random start solution along the most promising randomly explored paths. Genetic algorithms take a different approach by generating multiple random solutions and exploring the offspring solutions of the pair of best performing solutions. Lastly particle swarm optimization is a combination of the previous two, it uses multiple random solutions but instead of generating offsprings it moves the solutions along the most promising randomly explored paths. Each of these algorithms have their own advantages and disadvantages. The main reason why genetic algorithm were used for this thesis is due to their popularity in traffic modelling papers (Kesting & Treiber, 2008). This means that this algorithm has been thoroughly tested for this type of optimization problem and that there are multiple papers covering the best genetic algorithm parameters and objective functions (Kesting & Treiber, 2008; Ranjitkar et al., 2004).

8.2. Data

The first major point of discussion is the raw car trajectory data used to calibrate and test the various models. Both the UNINA and CARRS-Q dataset have flaws and possible shortcomings. Whether these flaws have an impact on the final conclusion is unclear but they do explain some of the anomalies in the results. Each of these flaws are discussed below.

The UNINA dataset has two flaws which stand out when the data is analysed.
First the mean headway of the drivers driving in the simulator setup is abnormally high. Where the headways of drivers in car-following, in most circumstances, is expected to be between 0.5 to 2.5 seconds (Saifuzzaman & Zheng, 2014) the average headway of the drivers in the UNINA dataset for scenario 4 is around 6 seconds and 10 seconds for scenario 5. That said this result is directly from the unprocessed data of the scenario. When the data are filtered to only include the car-following sections the headway drops to 5 seconds for both scenarios. This is still far above the expected value. As for the impact on the results and calibration of the model, the model should still calibrate properly since the trajectories are mostly composed of car-following data. The area where this anomaly does have an impact is on the calibrated model parameters. Instead of the more usual 1 second headway (T) and 5 meter standstill distance (s_0) (Saifuzzaman & Zheng, 2014) the model parameters will likely be much higher. This is exactly what was found in the calibrated model parameters with headways around 3 seconds for scenario 4 and 4 seconds for scenario 5 and a standstill distance of roughly 20 meters for scenario 4 and 10 meters for scenario 5. This is a bit lower than the average headway calculated from the data but when the standstill distance is converted to time by dividing it by the cars speed it translates in a total headway of 3.6 seconds for scenario 4 and 4.3 seconds for scenario 5. These large headways could also positively skew the error measurements compared to other papers. Since the errors are a relative metric, for the same absolute headway error the impact on the error is much smaller between this dataset and a dataset with a more ”normal” headway. So in short, the high headways in the data cause unusual calibrated parameters values, these parameters don’t have an impact for the results in the thesis itself but it makes it difficult to compare the error values of this thesis with other papers.
The second possible flaw is the correlation between task/driving difficulty and effort. The data explains that the effort measurement is based on these other measurements, namely the heart rate, pupil diameter, eyelid aperture, EEG workload which are all proxies for workload or effort. The i4Driving project and its related papers make the assumption that there is a strong correlation between the effort measurement and task difficulty and provide supporting evidence for this assumption (Irvine et al., 2023; Tang et al., 2023). However from the analysis done on the data of scenario 4 and 5 this correlation doesn’t always seem to be as strong or valid in all circumstances. When analysing the effort in scenario 4 and 5 it appears that the average effort is equal in both scenarios. If the effort and task difficulty are closely correlated this would be strange but this could indeed be the case if drivers lowered the TD for themselves, for example by keeping longer headways which does indeed happen in scenario 5. That said while keeping a longer headway would lower the average task difficulty it wouldn’t impact the peaks in task difficulty caused when a driver responds to one of the distraction questions. Yet these peaks are not observable in the graphs which include the effort, headway and gaze direction, for example see figure 8.1. With this as supporting evidence one of the following conclusions can be drawn, either the correlation between effort and task difficulty is not always applicable in all circumstances or the distractions experienced in this simulated experiment didn’t have a big impact on the task difficulty of the driver. In order to find which statement is true it is recommended that the experiment is performed again but this time with a stronger or mandatory, instead of optional, distraction. This should provide enough data to determine which of the two statements is correct.

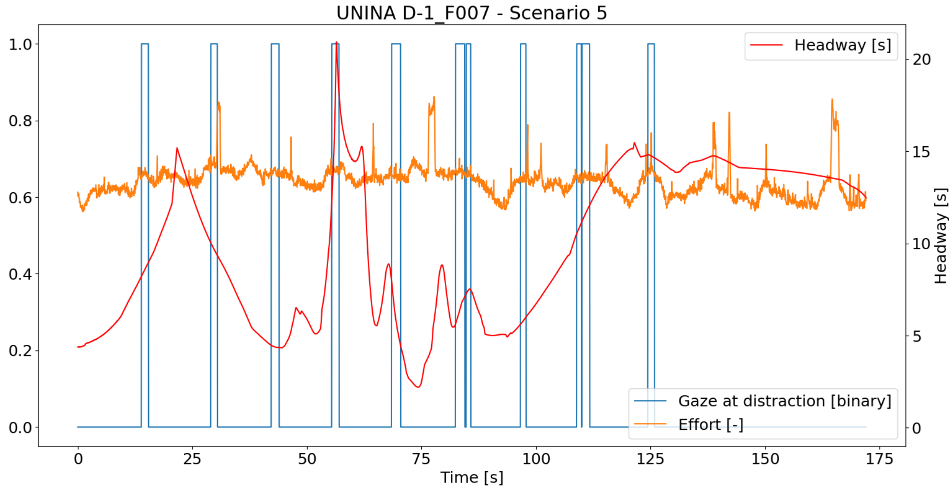

Figure 8.1: Effort, headway and distraction plot of driver F007 scenario 5.

Like the UNINA dataset the CARRS-Q dataset also has its flaws though these flaws are more design flaws than abnormalities in the results.
In total there are three minor issues which can all be attributed to manners in which the experiment is designed. First is the length of the trajectories. Since the experiment was designed for another study about driver interactions with pedestrians there are only relatively few and short sections where the dataset offers pure car-follow trajectory data. Because of this the duration of each trajectory is only between 20 and 30 seconds which is rather short compared to the UNINA data trajectories of 120 seconds. This might lead to some difficulties when calibrating the models and result in parameters which don't properly reflect the characteristics of the drivers. Another issue with the dataset is that the participants aren't representative of the mean population of drivers on the road. The main issue is that nearly all the participants are young drivers, younger than 25 with limited driving experience. This, again, has an impact on the average calibrated parameters, making it biased towards the behaviour of young inexperienced drivers. These short trajectories and uniform participant group both affect the same part, the parameters, because of this if there are abnormalities found in the parameters it is difficult to pinpoint its exact cause. That said, unless young drivers have vastly different behaviour to phone calls, these biases shouldn't have a big impact on the final conclusion to determine which model performs best with which distraction since all models are trained on the same biased data. The last design flaw with the CARRS-Q dataset is the speed of the leader. Halfway through the recorded trajectory the leader increases the speed. This isn't an issue in itself and in fact can be quite useful when calibrating some of the parameters. The issue is that this increase in speed is instantaneous providing no time windows where the participant can gradually increase his own speed. This could cause some issues for the car-following models; since the used car-following models have no concept of momentum they will likely follow with an instantaneous increase of acceleration which doesn't match the real trajectory of the drivers. While all the used models are based on IDM they do have some slight differences which could cause one of the models to perform better because of this particular abnormal situation. This increase in performance can be seen as an unfair bias since this is a situation which would never occur in the real world. Ideally this increase in performance is excluded when determining which model performs best but it's impossible to know which or if a model benefits from this illogical situation. To sum up, while the chances are low the conclusions of this thesis could be influenced by a design flaw in the leader speed and it is difficult to identify if this has indeed occurred. To test if this is the case, a small simulator experiment could be setup, with a similar leader speed jump but no other changes, to see if any model handles these jumps better than others. Or ideally the experiment could be redesigned and slightly modified to solve all three design flaws.

8.3. Results

The last section of the discussion covers the results of the individual trajectory calibrations and the calibrated models. While most results are as expected or have a clear explanation for why they deviated from the expected case, some of the results require a more in-depth analysis for why they deviated from their expected performance. This last category of results are analysed in this section with possible reasons for why they are unusual or recommendations on how to improve or verify these hypotheses.

8.3.1. Individual Trajectory Calibration

Similarly to the results chapter the first part covers the calibration process on the individual trajectories. There are two important points which can be identified as abnormal without a clear explanation.

The first abnormal point is the increased loss in performance in regards to headway by the Distraction model when it is used as calibrated model compared to the other models. As was pointed out in the results chapter, the other models lose around 10% performance in headway while the Distraction model loses around 15%. The most likely explanation, for this loss in performance when moving from an individual trajectory model to a calibrated model, is that the individual Distraction models were overfit on the individual trajectories. So the real question is not why did the Distraction models have a larger loss in performance but why were the Distraction models more overfit than the baseline models? There could be a couple of different reasons for this, either the calibration process behaved differently for the Distraction model or the issue lies with the parameter and internal logic for the Distraction model. Naturally the Distraction model used the exact same calibration process and number of total generations as the other models. The only difference is that the other simpler models with fewer parameters plateaued earlier and the calibration process was cut-off before reaching the total number of generations due to a lack of improvement. This usually didn't happen for the Distraction model, while its progress in calibration stalled around the same time as the other models it usually was able to make minor improvement until it reached the limit of total generations. This could be the source of the additional over-fitting but it could also be a symptom of a more fundamental problem. This leads to the second possible problem point, the internal logic and parameters of the Distraction model. The loss in performance when moving from an individual trajectory model to a calibrated model comes from the fact that each parameter now needs to represent a group instead of an individual. This is fine for more physical parameters which tend to be quite similar from one driver to another on the same road but for human factor parameters which are closely tied to the driving style and personality of the driver it is much harder to find a singular value which represents a group. The Distraction model is the model with the most human factor parameters, 5 in total, compared to 3 and 2 from the TDIDM and Multi-scale model. Furthermore the Distraction model has by far the most logic and equations relating to human factor parameters and variables. This allows the model to better express how a driver will react in certain situations but also makes it so that the model has the most to lose when averaging the parameters. That said according to this logic the model with the smallest loss in performance should be the IDM model but this is not confirmed by the data. The order of the models with the greatest loss to smallest loss is Multi-scale, IDM, distraction, TDIDM. For both datasets the IDM and Multi-scale are quite close to each other but the loss of the Multi-scale model is just a bit smaller than the IDM model hence contradicting the previous hypotheses. Because of this no definitive conclusion can be drawn for the question of why the Distraction model has such a big loss compared to the other baseline models. To answer this question it is recommended that a more in-depth analysis is performed to find which elements of models or data influence the loss in performance.

The second important point which was unusual in the individual calibrated trajectory results is that the KPI gap between the Multi-scale Five model and the baseline models reduces between the two UNINA scenarios but not for the CARRS-Q scenarios. This reduction in the gap in the UNINA scenarios is mainly caused by the fact that the Multi-scale Five models KPI reduced while the baseline models KPI stayed roughly the same or even increased slightly. The reduction in KPI from the Multi-scale Five model is expected since the scenario with the distraction is expected to be more chaotic and harder to predict. The illogical part is that the baseline models weren't affected by this harder to predict scenario. One possible explanation for this is as follows. A driver's behaviour is composed of two sections, a rational part which is predictable and can be modelled, and an irrational or chaotic part which can not

be modelled. The limit between the two is what the Multi-scale Five model tries to estimate. The hypothesis is that this boundary is lower for the Distraction model which causes the worse performance for the Multi-scale Five but not lower enough to start affecting the performance of the baseline models. See figure 8.2 for a visual aid regarding this hypothesis. This also provides a possible explanation for why the reduction didn't occur with the CARRS-Q dataset. The explanation would be that the Multi-scale Five model underestimated the boundary and that the subsequent reduction didn't affect it like the baseline models.
Regardless of whether this hypothesis is true or not the Multi-scale Five model isn't used in the final evaluation and its results do not affect the conclusion, its only purpose is to provide an estimate of how well models could perform. Because of this the topic isn't further studied in this thesis but since it is an interesting topic it is recommended that future research look into this topic and try to prove or disprove the above theory.

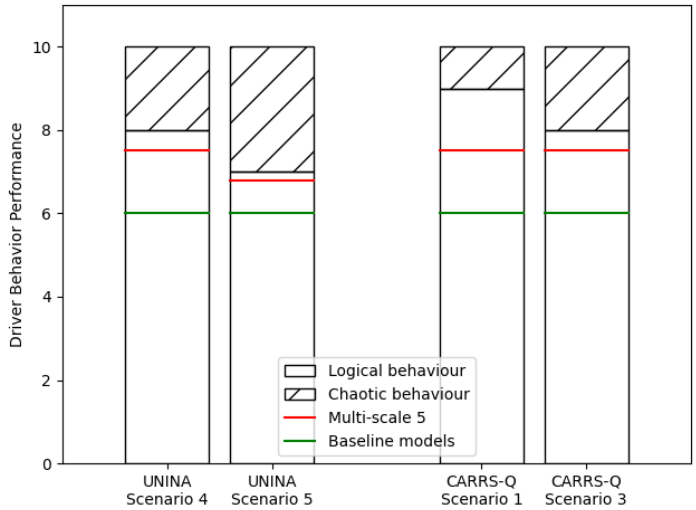

Figure 8.2: A visual aid representing the performance of various models in different conditions. Note all values are not representative of the real results.

8.3.2. Calibrated Model

The individual trajectory results aren't the only section in the results which deviated from the expected case. The calibrated model results also contain some abnormal situations. Four of these points are discussed in this subsection.

The first point of interest is that the Multi-scale Five model isn't included in the results. This exclusion was a deliberate choice. The reason for this is that the Multi-scale Five model serves a specific purpose in the individual trajectory analysis as theoretical limit of a model's performance. The Multi-scale Five model is built in such a way that it has access to more information than the other models and it makes an assumption which couldn't be made in an uncontrolled environment. This assumption relies on the model being used to predict a singular driver's behaviour, so trying to average the parameters and make a calibrated model would contradict this assumption. Naturally since this model isn't included in the evaluation the lack of corresponding calibrated model results doesn't impact the final conclusions.

The second unexpected result is the overall performance of the TDIDM model. From the results it's clear to see that the TDIDM performs the worst out of all the models in nearly every category. These results don't match the results and conclusions of the original paper by Saifuzzaman et al. (2015). According to the original paper TDIDM should perform better than IDM in every scenario, baseline and distracted, and every configuration, with or without human factor calibration. Naturally these contradicting results are worrying since they were performed in a nearly identical fashion on the exact same dataset. As a result of these findings a more thorough analysis on the differences in calibration methods was conducted. There are a total of three differences between the original paper and this thesis. First the KPI used is different, this paper uses the SMAPE and the original paper used the RMSPE. These

two error measurements are quite similar as they are both relative measurements but they handle big errors differently. So while these two KPIs can't be directly compared to one another their conclusions are still the same. For example if by using SMAPE it is found that IDM is significantly better than TDIDM, then RMSPE should find a similar conclusion and not the opposite conclusion. The second big difference is the order of calibration. In this thesis all parameters of the various models are calibrated in the same run. The original paper does this slightly differently, it calibrates the generic parameters using the baseline scenario and only calibrates the HF parameters using the distracted scenario. While the second option can seem advantageous at first glance it offers no real benefits other than reduced computational complexity. After all by calibrating every parameter at the same time it doesn't limit the interactions between the different parameters during calibration so it should be able to find the same parameter combination as the second method. Lastly the final difference between the two calibration methods is the software used to calibrate the models. In both cases a genetic algorithm is used; this thesis uses the PyGAD package for Python and the original paper used the built-in GA from Matlab. Both tools have been used by countless people and should converge to nearly identical results.
To more accurately compare the results of the two tests the results of this paper were rerun using the same methods used in the original paper. This means using the RMPSE of the spacing and calibrating in two steps. The only part which remains different is the software tool used for the genetic algorithm. The results of this can be found below. It is worth noting that this additional test is only run for the CARRS-Q scenario 1, which corresponds to validation 1 in the original paper, and that the results of the thesis validation runs are the average of the three seeds. Sadly no t-test can be performed between the original paper and thesis results to see if a method is significantly better than the other since the standard deviation of the results is not given in the original paper.

	IDM	TDIDM
TDIDM paper (validation 1)	24.41%	19.06%
Thesis one step calibration	10.3%	16.8%
Thesis two step calibration	N/A	16.5%

Table 8.1: Spacing RMSPE for scenario 1 of the CARRS-Q dataset validation results.

The additional test shows that there is still a big difference between the results of the original paper and the thesis. This difference is mainly in the paper results of the IDM model, it seems that the IDM model performed far worse than it should in the paper. This is still true when comparing it to the worst seed of the IDM model in the thesis, which had an error of 13.1%. To give a more complete and accurate conclusion it would be preferable to repeat this test in Matlab but that falls outside the scope of this thesis.

The third result which deviates from expected behaviour is the calibrated desired speed of the Multi-scale model in the CARRS-Q scenarios. The simulator experiment design provides the speed of the leading vehicle during the simulation, starting at 20 km/h and increasing towards 35 km/h when the spacing conditions are met. The found desired speed parameter of the IDM model is around 45 km/h which is higher than the leading vehicle speeds. That said such a finding is inline with findings of other papers regarding the desired speed of vehicle in low speed traffic (Saifuzzaman & Zheng, 2014), since it also needs to accommodate for the high speed peaks of the driver. The issue with the desired speed of the Multi-scale model is that not only it is much higher than the leading vehicle speed, it is also much higher than the peak participants speed at around 90 to 120 km/h. This is a strange result and without additional information it looks like it isn't the optimal solution. That said when paired with information about the task saturation and speed reduction caused by being overloaded it seems much more reasonable. As can be seen in figure 7.6, the driver experiences a high task demand during the second half of the simulation. This situation is not only applicable to this particular driver but happens to nearly all drivers with some already experiencing some high task demand peaks during the first half of the simulation. Due to this a desired speed reduction is applied by the model which range from -40% to -60% in some cases. This means that the applied desired speed during simulation is between 36 to 72 km/h which nicely overlaps with the speed peaks in the second half of the simulation. As for why the genetic algorithm found this to be the optimal solution is hard to say, after all picking a lower base

desired speed paired with a lower distraction task difficulty should have the same effect on the model. This means that either the calibration process got stuck in a local minima or the combination of high desired speed with high distraction task demand provides an additional (minimal) benefit which is difficult to discover in the trajectories.
Another possible explanation for this extremely high desired speed is that the driver's true desired speed can't be derived from pure car-following data. The reason for this is that a driver in car-following mode is, per definition, limited by the leader. This means that their driving speed is lower than, or coincidentally at, their true desired speed. So the observed driving speed isn't the ego's desired speed but the leaders desired speed. This also means that the drivers desired speed can be estimated to be much larger than in reality without negative consequence hence the large differences in calibrated desired speed for the CARRS-Q dataset. This explanation provides a feasible answer to both scenarios in the CARRS-Q dataset whereas the previous explanation only provides an answer for the distracted scenario. A counterargument for this explanation is that these large variances in desired speed are not observed in the UNINA calibration results nor in the different seeds.
In terms of impact on the conclusion the effect should be relatively small. In the case that the GA is stuck in a local minima it would mean that the performance of the model is reduced but is still close to the real optimal solution. This is due to the properties of genetic algorithms which make them quite successful at evading shallow local minima. And if the high desired speed is instead the optimal solution that would mean that the performance of the model is increased, leading to a more positive conclusion.

The last abnormal result with no clear explanation is the task demand of the spontaneous distraction in CARRS-Q scenario 3. The initial hypothesis for the distraction task demand was that it would spike at the sudden phone call and that it would be reduced during the actual conversation. For most drivers this doesn't seem the case and instead the opposite happens; the spontaneous phone call results in a relatively low task demand and the conversation a much higher task demand. This counter intuitive example does seem to have an explanation when analysing the speed plots of the drivers, see figure 7.6 as example. Most drivers seem to have a reaction which can be split into two parts. When the driver first receives the phone call they decelerate, in the example it happens quite aggressively and the driver loses around 6 km/h in roughly 2 seconds. After this initial deceleration the driver tends to accelerate quite aggressively over a slightly longer period of time and they usually also overshoot their previous speed. It's during this re-acceleration or shortly after that the phone call is picked up. This means that the model needs to replicate two different behaviours during the spontaneous phone call interval using a single variable, the spontaneous task demand. This is an issue since the Multi-scale model is designed to react in a single fashion when experiencing a high task demand, namely reduce the desired speed when experiencing a high task demand. This means that the model can only match one of the two parts of the reaction and since the acceleration section seems to last a bit longer than the deceleration section it also means that the acceleration section holds more weight in the KPI calculation. So when the model is forced to choose between applying a low or high task demand, it chooses the low task demand to match the strong acceleration.
Forcing the Distraction model to choose between the two reactions causes it to miss out on overall performance hence it has a negative impact on the final conclusion of the Distraction model. Currently it is not known whether this more complex reaction is associated with spontaneous distractions or if it is something specific to this dataset. It is recommended that future research look into the reaction to spontaneous distractions and uses that understanding to build a more performant Distraction model.

9
Conclusion

The goal of this thesis was to fill a gap in the existing literature about the implementation of various distractions in task-capability interface traffic models. This objective is concluded in this chapter whose purpose is to summarise all the found results and analyses in short concise answers.
To this end this chapter is divided in two sections. The first provides answers to the research questions posed in chapter 2 Research Question. The sub-questions are answered first and the main research question is answered last. The second section addresses the recommendations for future works and ways of using the found results of this thesis.

9.1. Answer to the Research Questions

To fulfil the goal and fill the gap defined in the introduction a main research question was formulated. This question is as follows: *"How can different types of distractions be implemented and validated at a microscopic scale in a task-capability interface car-following model?"*. The main question is then broken down in four sub-question. These questions are answered in chronological order below using the found results. This will lead to the answer of the main question.

What should the distraction framework properties be in order to support different types of distractions?
The purpose of this first sub-question is to properly define what distraction are and how they should be modelled. Since literature can have varying definitions of what does and doesn't count as a distraction and how they interact with drivers, it is important that this is properly defined in this thesis. The answer to this question is primarily based on the findings of the literature review which are presented in chapter 4 Distraction Framework Definition and is as follows. The framework used to model distractions will rely on the low-level characteristics of distractions. These characteristics allow to more easily model complex distractions without completely modelling each distraction individually from the ground up. Furthermore three distraction lifecycle stages were identified which are each influenced by the low-level characteristics. These stages are the distraction trigger, the distraction intensity and the distraction effect. These three stages suffice to determine who, how much and in which manner the distraction influence drivers.

Which TCI model implementation should be used as a base model to implement the distraction framework?
Due to the scoping of the thesis it is unnecessary to build a car-follow model from the ground up. Instead an existing TCI car-following model implementation can be chosen and modified in order to implement the distraction framework. To this end a short multi-criteria analysis was performed in chapter 5 TCI Multi-criteria Analysis. This led to the selection of the Multi-scale model by Calvert et al. (2020) and Van Lint and Calvert (2018) as the most promising model. In particular this model offers great flexibility with its modular design while keeping adequate implementations opportunities for other human factors without greatly increasing model complexity.

How should the logic and parameters of the base TCI model implementation be modified in order to integrate the distraction framework?
To validate the distraction framework it needs to be implemented and tested in the chosen TCI car-following model. This implementation should have minimal effects on the existing functions of the chosen TCI model while still providing the full benefits of the distraction framework. This work has been done in chapter 6.2 Distraction Model Implementation. Due to its modular nature, the Multi-scale model provides multiple entry points for the distraction framework to use and modify. The distraction trigger and distraction intensity stages hook into the environment to task difficulty calculations, thus adding a more nuanced control for distractions. The distraction effects provides additional methods of influencing the driver stimuli and preferences using either one of the global task saturation variables or a specific distraction intensity. This integration of distraction framework and Multi-scale model is called the Distraction model.

Which distraction types and TCI model implementation should be used to validate the proof of concept for the distraction framework?
The level and quality of validation of a new design depend heavily on the methods and data used in the validation tests. To this end it's important to find appropriate components for the test. For the validation of the distraction framework no new methods or data was gathered, instead existing proven methods were used and proper datasets were found with needed overlapping characteristics. These findings are presented in 6 Validation Methodology. The two selected datasets provided trajectories containing vastly different distractions, a continuous mental-visual distraction and a spontaneous auditory distraction. Additionally the baseline evaluations of these trajectory are done with the IDM, TDIDM and Multi-scale models. The method used to calibrate the newly built Distraction model is a genetic algorithm which optimizes for headway.

How can a framework for systematically defining different types of distractions be implemented and validated at a microscopic scale in a task-capability interface car-following model?
With all four sub-questions answered the main research question can now be answered.

The manner in which a framework which systematically defines distractions can be implemented and validated in a car-following model is by following the steps outlined by the individual sub-questions. This means first defining the logic and properties of the distraction framework. Subsequently an adequate TCI car-following model needs to be chosen and modified by implementing the framework. Lastly this Distraction model needs to be tested to prove that it is valid for the given distraction scope.
The thesis in itself follows this guide and the results of the validation test for the built Distraction model show that these steps work and that the distraction framework provides adequate support to systematically model different types of distractions. To be more specific the results of the validation showed that the built Distraction model provided equivalent results to other models, which can support distractions and human factors, for at least two completely different types of distractions. This also means that a systematic approach of modelling distraction won't necessarily lead to an increase in performance for the chosen model. Instead it makes the model more flexible, easier to understand and simpler to expand upon.

With the main research question answered it provides a solution to fill the gap in literature. Other alternative solution for filling this gap undoubtedly exist and should be studied in future research.

9.2. Future Developments

As mentioned at the end of the last section the gap in existing literature has been filled by one solution. This is the minimum amount of solutions necessary to fill this gap. Other manners of filling this gap are possible and might provide interesting pros and cons compared to the found method. To this end it is recommended that others research this topic of systematically modelling distractions in traffic models. This can be done either by improving upon this thesis or by taking a completely new approach. In the case of wanting to continue improving upon this thesis, the chapter 8 Discussion & Recommendation contains multiple suggestions and weak points of this thesis. By strengthening these weaknesses it should be possible to obtain better results. One recommendation which isn't mentioned in the discussion is that the amount of validated building blocks/equations for the distraction trigger, intensity and effects are quite limited. This is due to the limited amount of distraction datasets available. By increasing the amount of building blocks it will become easier for others to use this framework for their own separate studies.

This work can also be used to further studies into driver behaviour when encountering distractions. This systematic approach to defining and modelling distractions should help to more systematically understand and analyse the behaviour of drivers.

A

MCA Model Evaluation

Based on the criterion explained in chapter 5.1 MCA Criteria a more in depth-analysis of the available TCI models is performed. There are a total of three models, which will be evaluated in this section. Each of these models has an overview of their purpose and functionally in the literature review, see chapter 3.2 TCI Model Implementation.

A.1. Multi-scale Framework

The first model which will be looked at is the Multi-scale framework by Calvert et al. (2020) and Van Lint and Calvert (2018), a more detailed explanation of the model can be found in chapter 3.2.1 Multi-scale Framework.

First is the evaluation of the model complexity and flexibility. The Multi-scale model is designed to be a modular model at its core. There are two primary layers, the cognition layer and the control layer. These two layers can be completely separated and changed for another layer with a similar function as long as it respects the inputs and outputs of the layers. This design concept also applies to the components inside the primary layers. This makes the model extremely flexible, after all the modules can be calibrated on a modular level and changing one module shouldn't have a large impact on the parameter's calibration of the other modules. Additionally, this modular design has allowed the authors of the model to increase the number of intermediate steps and number of calculations to be executed in a singular time-step without losing oversight of what the model is doing. This results in a model which is complex but still understandable. So in terms of model flexibility the model scores quite high while for complexity it scores middle high.

In terms of computational complexity, the model is surprisingly simple. While there are quite a few steps to go through, the steps themselves are mostly basic additions and multiplications. On occasion it also evaluates a simple function but other than that no computationally heavy mathematics. Furthermore there are no direct feedback loops inside a singular time step but time steps do need some data from previous time steps which increases the computational complexity slightly. Hence a middle high score for computational complexity.

In third is the evaluation of the model support for human factors and distractions. Since the Multi-scale model is a TCI model it has an implementation of the task difficulty. This task difficulty is one method of representing the human factor elements of drivers. That said since all the models are TCI model, the parts that make them unique is how this task difficulty is calculated and what it affects. The way it's calculated is with a modular system where the task difficulty is the sum of all the underlying tasks, these tasks can both be primary tasks and secondary tasks making it quite flexible. Usually the state of these tasks are expressed with a single variable which is then transformed into task difficulty but that isn't strictly required. Instead a task could calculate its difficulty with another type of multivariable function. The effects of the task difficulty are also quite broad. The summed task difficulty can influence the reaction time and the various driver preferences (e.g. desired headway). Having said that

a singular task can't have a direct effect on the driver, instead it's first bundled together which limits the specific effect of a singular task on a driver.
In terms of distractions, this modular system allows the user to implement distractions quite easily by adding a new temporary task. By default the model doesn't provide any distractions tasks but since the range of inputs from the control layers are quite diverse it's possible to implement a large range of distractions. That said some more specific and unusual distractions might run into issues with the provided inputs. Since this model wasn't built to accommodate a specific type of distraction it doesn't account for the more exotic inputs needed for some distractions (e.g. ocular occlusion or eye tracking).

Lastly in terms of personal preferences criteria the model scores full marks. The author is easy to contact since it's the chair of this thesis. Furthermore the model and its development is fully open source, the code implementation can be found on GitHub[1]. Lastly as bonus points the model is implemented in open traffic sim (OTS), which is a modular traffic simulation software developed at TU Delft, and it's the traffic simulation software I'm most familiar with.

A.2. Task Difficulty Car-following

The second model is the Task Difficulty Car-Following model (TDCF) model by Saifuzzaman et al. (2015, 2017), a more detail explanation on how this model works can be found in chapter 3.2.2 Task Difficulty Car-Following. This model designs a task difficult formula and uses the resulting value to influence existing car-following models. The author does this with two well-known models, the IDM model and the Gipps model, resulting in the TD-IDM and TD-Gipps model.

In terms of model complexity this model is rather simple. There is one formula to calculate the task difficulty which is based on headway with a few extra driver preference factors and a slight modification to either the desired speed or desired headway in the model control logic. The new parameters are also quite self-explanatory except for the task difficulty sensitivity parameter which has little to no effect on the driver's behaviour in most cases. This lack of steps and simplicity makes it so that the model exhibits less flexibility. While it's relatively straightforward to implement this logic in an existing model, modifying it to incorporate a more advanced distraction system will require a complete recalibration of all the task difficulty parameters.

In terms of computational complexity, the model is about as light as can be. Both the IDM and Gipps model are known to be lightweight and performant models and the new modifications barely add any extra computational load to this.

In terms of human factors this model does add other parameters in addition to the task difficulty. Namely the risk parameter and the modified reaction time. The risk parameter is a representation of how aggressive a driver is when faced with reduced driving capabilities. This parameter allows for aggressive, neutral and mild drivers compared to the original base models. While this implementation doesn't have any of the more nuanced human factors like social pressure it is a good surrogate for implementing risk aversion which leads to more accurate human driving behaviour. The modified reaction time is the model's implementation of distractions. It's a driver specific parameter which increases the reaction time of the driver when in a distracted scenario. Since the modified reaction time is a fixed parameter, it doesn't allow for more dynamic distraction nor different levels of distractions. It's fixed at one level of distraction per driver per scenario. While this model does have the most important variables needed in order to simulate the most common distractions this lack of distraction effects makes it unsuitable for the implementation of more complex distractions.

As for the last two criteria, the author is a university professor so it shouldn't be too difficult to contact the author. The model code is also open source, that said the used data aren't available. The paper uses data from an experiment made for another paper which fits its needs but the original paper didn't publicly release its code or data.

[1] https://github.com/averbraeck/opentrafficsim

A.3. Fuzzy Task Difficulty

The fuzzy task difficulty longitudinal control model (FTD-LCM) by Li et al. (2020) is a model that merges multiple existing models theories in order to fill a gap in the literature. The two primary goals are to have a better representation of the human brains ambiguity and to use a more accurate base car-following model. For a more detailed explanation see chapter 3.2.3 Fuzzy Task Difficulty Longitudinal Control Model.

When trying to understand the FTD-LCM model it will become clear that the model doesn't have more steps than the models it's build upon. In fact it removes and condenses some of these steps into a singular step in order to keep the logic from getting too complex. The main issue with this model's complexity is that the steps themselves are not that clear. This is an inherent problem of fuzzy logic. Since fuzzy logic relies on a combination of probability distributions and concrete inference rules it can be difficult to follow what the model is doing in detail. On the other hand the car-following part of the model, the LCM model, is a model based on real human perception variables, making it more intuitive to understand than other car-following models. Contrary to this the model's flexibility is quite poor. Due to the nature of fuzzy logic inference rules it becomes quite difficult to expand upon them without having to completely rebuild the logic section. This is not a simple recalibration or modification to the logic but a complete rewrite.

In terms of computational complexity the model is a bit more complex than the typical IDM or Gipps models. The majority of the calculations are simple additions and multiplications. Having said that there are also a few non integer exponentials and most importantly probability distributions. While a few probability distributions evaluations aren't problematic, the model requires each vehicle to poll a result multiple times in each time step. This adds up to quite a few pseudo random number generations per time step slowing down the simulation overall.

As for the fourth and fifth criteria in the evaluation, the model does quite well when it comes to human factors. At its core the LCM model tried to emulate the comfort zone of the driver, and by keeping other cars out of its comfort zone it results in human like driving behaviour. This is then further enhanced by the perception and personal preference biases introduced by the reduced situational awareness from task difficulty. Though the model doesn't offer much control over the variance of individual drivers, the nature of fuzzy mathematics makes it so that each driver will behave slightly differently. As for distractions the model doesn't offer any built-in distraction mechanisms. While distractions could be incorporated into the task difficulty the model makes no extra modifications to the original models it's based on and only tests a flat increase in task difficulty. This isn't inherently bad, but the concept of fuzzy distractions hasn't been explored. As for adding new distractions, the necessary inputs and outputs still exist though some specific distractions might need more data to be added to the model.

The author of the paper should be fairly easy to contact since he's a professor at a university. In terms of open source and ease of access, none of the model code nor the data is publicly available. Furthermore, some versions of the paper are also behind paid services.

B
Model Code

This chapter gives a brief overview of the model code. The project code can be found at https://github.com/DwarfyAssassin/Distraction-framework. This repository only contains the project code, not the data nor the results.

The following table is a short overview of the main code files:

File name	Description
`main_calibrate.py`	The main file
`car.py`	Contains the car class
`data_loader.py`	Helper functions to load and format data files
`helper.py`	Helper functions, including KPI
`model.py`	Contains the various model logic and classes
`optimize.py`	Contains the main optimization logic and parameter bounds
`results.py`	Manages the results files
`scenario.py`	Abstract scenario class
`scenario_TDIDM.py`	CARRS-Q scenario class
`scenario_UNINA.py`	UNINA scenario class
`scenario_synthetic.py`	Synthetic scenario class

C

Model Formulas

This appendix provides the exact formulas and constants used for the various external models in this thesis.

C.1. IDM

$$a(t) = max\left[-b_{max},\ a_{max}\left(1-\left(\frac{v(t)}{v_0}\right)^{\beta}-\left(\frac{S^*(t)}{S(t)}\right)^2\right)\right] \tag{C.1}$$

$$S^*(t) = s_0 + v(t)T - \frac{v(t)\Delta v(t)}{2\sqrt{a_{max}b_{comf}}} \tag{C.2}$$

These formulas use the following constants:
$\beta = 4$

C.2. TDIDM

These equations are taken from Saifuzzaman et al. (2015, 2017)

$$a(t+\tau') = max\left[-b_{max},\ a_{max}\left(1-\left(\frac{v(t)}{v_0}\right)^{\beta}-\left(\frac{S^*(t)*TD(t)}{S(t)}\right)^2\right)\right] \tag{C.3}$$

$$S^*(t) = s_0 + v(t)T - \frac{v(t)\Delta v(t)}{2\sqrt{a_{max}b_{comf}}} \tag{C.4}$$

$$TD(t) = \left(\frac{v(t)T}{(1-\delta)S(t)}\right)^{\gamma} \tag{C.5}$$

$$\tau' = \tau + \varphi \tag{C.6}$$

These formulas use the following constants:
$\beta = 4$

C.3. Multi-Scale

These equations are taken from Calvert et al. (2020) and Van Lint and Calvert (2018)

$$a(t+\tau'(t)) = max\left[-b_{max},\ a_{max} * min\left[1-\left(\frac{v(t)}{v_0'(t)}\right)^{\beta},\ 1-\left(\frac{S^*(t)}{S'(t)}\right)^2\right]\right] \tag{C.7}$$

$$S^*(t) = s_0 + v(t)T'(t) - \frac{v(t)\Delta v'(t)}{2\sqrt{a_{max}b_{comf}}} \tag{C.8}$$

$$T'(t) = (1 - \beta_T * \epsilon_{TS}(t))T \tag{C.9}$$

$$v'_0(t) = (1 - \beta_{v_0} * \epsilon_{TS}(t))v_0 \tag{C.10}$$

$$\tau'(t) = \epsilon_{SA}(t) * \tau_{a,\ max} + \tau \tag{C.11}$$

$$\Delta v'(t) = (1 + \delta_{sign} * \epsilon_{SA}(t))\Delta v(t) \tag{C.12}$$

$$S'(t) = (1 + \delta_{sign} * \epsilon_{SA}(t))S(t) \tag{C.13}$$

$$\epsilon_{SA}(t) = SA_{max} - SA(t) \tag{C.14}$$

$$\epsilon_{TS}(t) = clamp[0,\ TS(t) - TS_{crit},\ 1] \tag{C.15}$$

$$SA(t) = \begin{cases} SA_{max} & TS(t) < TS_{crit} \\ SA_{max} - \frac{TS(t) - TS_{crit}}{TS_{max} - TS_{crit}} * (SA_{max} - SA_{min}) & TS(t) >= TS_{crit}\ \&\ TS(t) < TS_{max} \\ SA_{min} & TS(t) >= TS_{max} \end{cases} \tag{C.16}$$

$$TS(t) = \frac{TD_{total}(t)}{TC} \tag{C.17}$$

$$TD_{total}(t) = TD_{cf}(t) + TD_d(t) \tag{C.18}$$

$$TD_{cf}(t) = \begin{cases} TD_{cf,max} & h(t) <= h_{min} \\ TD_{cf,0} + \frac{h(t) - h_0}{h_{min} - h_0} * (TD_{cf,max} - TD_{cf,0}) & h_{min} <= h(t) <= h_0 \\ TD_{cf,0} & h(t) > h_0 \end{cases} \tag{C.19}$$

$$h_{min}(t) = \begin{cases} \left(1 + \frac{a(t) + b_{comf}}{b_{max} - b_{comf}}\right) h_{min} & a(t-1) < -b_{comf} \\ h_{min} & else \end{cases} \tag{C.20}$$

These formulas use the following constants:
$\beta = 4$
$\beta_T = 0.9$
$\beta_{v_0} = 0.9$
$\tau_{a,\ max} = 2$
$SA_{max} = 1$
$SA_{min} = 0.5$
$TS_{crit} = 0.8$
$TS_{max} = 2$
$TC = 1$
$TD_0 = 0.5$
$TD_{cf,max} = 1$
$h_0 = 3s$
$h_{min} = 1s$

D

Glossary

This appendix contains a glossary of all used acronyms and variables in the thesis. They are sorted by category for easier perusal. (See next page)

Abbreviation	Term	Definition
General Car-following Model Formula Variables		
s_i**	Set of stimuli	The set of all available stimuli in a traffic model.
θ_i**	Set of driver preferences	The set of all driver preferences in a traffic model.
ω_i**	Set of world characteristics	The set of all world variables and parameters in a traffic model.
b_{max}	Maximum deceleration	The physical maximum deceleration of the vehicle.
a_{max}	Maximum acceleration	The physical maximum acceleration of the vehicle.
b_{comf}	Maximum comfortable deceleration	The maximum comfortable deceleration of a driver.
v_0	Desired speed	The comfortable driving speed of a driver in a given environment.
s_0	Desired standstill distance	The comfortable distance between a car and their leader at standstill.
T	Desired headway	The comfortable time gap between a car and their leader.
τ	Reaction time	The delay between perception and action.
TCI Model Formula Variables		
TC_i	Task capacity	The total mental capacity of a driver.
$TD_{cf,i}$	Car-following task demand	The mental demand generated by the car-following task.
$TD_{d,i}$	Distraction task demand	The demand generated by the distraction task.
$TD_{total,i}$	Total task demand	To sum of all task demand.
TS_j	Task saturation	The ratio between total task demand and task capacity.
TD_i**	Task difficulty	Same as task saturation.
SA_i	Situational awareness	The awareness level of a driver.
δ_{sign}	Perception bias direction*	Determines whether a driver increases or decreases perceived stimuli.
AR_i	Anticipation reliance	The reduction factor for secondary task demand due to anticipation.
δ	Risk parameter	The risk assessment and response behaviour of drivers.
γ	Sensitivity parameter	The perceived task difficulty exponential.
φ	Reaction time increase	The increase in reaction time during distracted situations.
Distraction Framework Formula Variables		
$DT_{i,j}$ / f_{DT}	Distraction trigger	The formula for the distraction trigger of a distraction.
$DI_{i,j}$ / f_{DI}	Distraction intensity	The formula for the distraction intensity of a distraction.
$DE_{i,j}$ / f_{DE}	Distraction effect	The formulas for the distraction effects of a distraction.
DS_j	Distraction strength	The inherent strength level of a distraction.
DDS_i	Driver distraction sensitivity	The drivers sensitivity towards a certain distraction type.
SP_j	Spontaneous parameter	Whether a distraction appears spontaneously.
$TD_{dsp,i}$	Spontaneous distraction task demand	The task demand associated with a spontaneous distraction.
$\beta_{a,i}$	Maximum acceleration reduction factor	The maximum acceleration reduction factor due to distractions.

Table D.1: *Informal name/no official name. **Only used in literature study.

Abbreviation	Term	Definition
	Traffic Models	
CF	Car-following	A longitudinal traffic model.
LC	Lane change	A lateral traffic model.
HF	Human factor	A part of human behaviour which falls outside typical logic.
OTS	Open traffic sim	A open-source traffic simulation program.
FTD-LCM	Fuzzy Task Difficulty Longitudinal control model	A type of TCI model.
TCI	Task-Capability Interface	A type of traffic model.
	Task-Capability Interface Model	
TD	Task Demand/difficulty	The mental load of a task.
TC	Task capacity/capability	The total mental capacity of a driver.
TS	Task saturation	The ratio between TD and TC.
SA	Situational awareness	The level of awareness of the driver of their surroundings.
	Distraction Terms	
NDRT	Non-driving task	A secondary task which doesn't contribute to safe driving.
DT	Distraction trigger	The moment when a distraction applied to a driver.
DI	Distraction intensity	The strength with which the distraction applies applies to a driver.
DS	Distraction strength	The base unmodified strength of a distraction.
DE	Distraction effect	The behavioural effects of a distraction.
DDS	Driver distraction sensitivity	A DS scaling factor for a driver.
	Other	
MCA	Multi criteria analysis	A structured approach to determine the preferred alternative.
KPI	Key performance indicator	A quantifiable measure which represents the most important aspect of an objective.
RMSE	Root mean squared error	A commonly used absolute error metric.
RMSPE	Root mean squared percentage error	A commonly used relative error metric.
SMAPE	Symmetric mean absolute percentage error	A relative error metric.
df	Degrees of freedom	A statistical representation of the variance in a dataset.
CDF	Cumulative density function	The sum of a probability density function.
GA	Genetic Algorithm	A type of optimization algorithm.
UNINA	University of Naples Federico II	A Italian university.
CARRS-Q	Centre for Accident Research and Road Safety	A research institute attached to the Queensland University of Technology.

Bibliography

Andersen, G. J., & Sauer, C. W. (2007). Optical information for car following: The driving by visual angle (dva) model. *Human Factors The Journal of the Human Factors and Ergonomics Society*, *49*(5), 878–896. https://doi.org/10.1518/001872007x230235

Armstrong, J. S. (1985, January 1). *Long-range forecasting: From crystal ball to computer* (2nd. ed.). Wiley-Interscience.
ISBN 978-0-471-82260-8.

Bevans, R. (2020, January 31). *An introduction to t tests*. Retrieved March 30, 2025, from https://www.scribbr.com/statistics/t-test/

Calvert, S. C., Schakel, W. J., & Van Lint, J. (2020). A generic multi-scale framework for microscopic traffic simulation part ii – anticipation reliance as compensation mechanism for potential task overload. *Transportation Research Part B Methodological*, *140*, 42–63. https://doi.org/10.1016/j.trb.2020.07.011

Cuentas-Hernandez, S., Li, X., King, M. J., & Oviedo-Trespalacios, O. (2023). The impact of road traffic context on secondary task engagement while driving. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1139373

Eltoweissy, M., Olariu, S., & Younis, M. (2010). Towards autonomous vehicular clouds: A position paper (invited paper). *49*, 1–16. https://doi.org/10.1007/978-3-642-17994-5_1

Engström, J., Monk, C. A., Hanowski, R. J., Horrey, W. J., Lee, J. D., McGehee, D. V., Regan, M., Stevens, A., Traube, E., Tuukkanen, M., Victor, T., & Yang, C. Y. D. (2013). *A conceptual framework and taxonomy for understanding and categorizing driver inattention* (No. 9984186943902771). European Commission. Directorate General for Communications Networks, Content & Technology. Retrieved February 15, 2025, from https://iro.uiowa.edu/esploro/outputs/report/A-conceptual-framework-and-taxonomy-for/9984186943902771

Flores, B. E. (1986). A pragmatic view of accuracy measurement in forecasting. *Omega*, *14*(2), 93–98. https://doi.org/10.1016/0305-0483(86)90013-7

Fuller, R., McHugh, C., & Pender, S. (2007). Task difficulty and risk in the determination of driver behaviour. *European Review of Applied Psychology*, *58*(1), 13–21. https://doi.org/10.1016/j.erap.2005.07.004

Fuller, R. (2011, January 1). *Driver control theory* (Porter, Ed.). Elsevier Inc. https://doi.org/10.1016/b978-0-12-381984-0.10002-5

Gordon. (2008, January 1). *Crash studies of driver distraction*. CRC Press. https://doi.org/10.1201/9781420007497.ch16

Hamdar, S. H., Mahmassani, H. S., & Treiber, M. (2015). From behavioral psychology to acceleration modeling: Calibration, validation, and exploration of drivers' cognitive and safety parameters in a risk-taking environment. *Transportation Research Part B Methodological*, *78*, 32–53. https://doi.org/10.1016/j.trb.2015.03.011

Hamdar, S. H., Treiber, M., Mahmassani, H. S., & Kesting, A. (2008). Modeling driver behavior as sequential risk-taking task. *Transportation Research Record Journal of the Transportation Research Board*, *2088*(1), 208–217. https://doi.org/10.3141/2088-22

Irvine, Zhang, Miranda, García, Rodríguez, Saco, Koyal, Rad, Rusciano, Stang, Montanino, & Punzo. (2023, July 27). *Methodology and results: Relevant use cases and safety-critical scenarios* (No. D1.6). https://i4driving.eu/deliverables/

Kesting, A., & Treiber, M. (2008). Calibrating car-following models by using trajectory data. *Transportation Research Record Journal of the Transportation Research Board*, *2088*(1), 148–156. https://doi.org/10.3141/2088-16

Kircher, K., Larsson, A., & Hultgren, J. A. (2013). Tactical driving behavior with different levels of automation. *IEEE Transactions on Intelligent Transportation Systems*, *15*(1), 158–167. https://doi.org/10.1109/tits.2013.2277725

Lewis-Evans, B., De Waard, D., & Brookhuis, K. A. (2010). That's close enough—a threshold effect of time headway on the experience of risk, task difficulty, effort, and comfort. *Accident Analysis & Prevention*, *42*(6), 1926–1933. https://doi.org/10.1016/j.aap.2010.05.014

Li, L., Li, Y., & Ni, D. (2020). Incorporating human factors into lcm using fuzzy tci model. *Transportmetrica B Transport Dynamics*, *9*(1), 198–218. https://doi.org/10.1080/21680566.2020.1837033

Liang, Y., & Lee, J. D. (2010). Combining cognitive and visual distraction: Less than the sum of its parts. *Accident Analysis & Prevention*, *42*(3), 881–890. https://doi.org/10.1016/j.aap.2009.05.001

Ni, D., Leonard, J. D., Jia, C., & Wang, J. (2015). Vehicle longitudinal control and traffic stream modeling. *Transportation Science*, *50*(3), 1016–1031. https://doi.org/10.1287/trsc.2015.0614

Olstam, Andersson, Ahlström, Kircher, Johansson, García, Dopico, Quinteiro, Galante, Fucito, Montanino, Lindner, Masceti, Finkeldei, & Zhang. (2023, December 19). *D3.2 experimental setup for the driving simulator experiments* (No. D3.2). Retrieved April 15, 2025, from https://i4driving.eu/deliverables/

Papathanasopoulou, V., & Antoniou, C. (2015). Towards data-driven car-following models. *Transportation Research Part C Emerging Technologies*, *55*, 496–509. https://doi.org/10.1016/j.trc.2015.02.016

Phan, N. M. T., Thouvenin, I., & Fremont, V. (2016). Enhancing the driver awareness of pedestrian using augmented reality cues. *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, 1298–1304. https://doi.org/10.1109/itsc.2016.7795724

Punzo, V., Ciuffo, B., & Montanino, M. (2012). Can results of car-following model calibration based on trajectory data be trusted? *Transportation Research Record Journal of the Transportation Research Board*, *2315*(1), 11–24. https://doi.org/10.3141/2315-02

Radhakrishnan, V., Louw, T., Gonçalves, R. C., Torrao, G., Lenné, M. G., & Merat, N. (2023). Using pupillometry and gaze-based metrics for understanding drivers' mental workload during automated driving. *Transportation Research Part F Traffic Psychology and Behaviour*, *94*, 254–267. https://doi.org/10.1016/j.trf.2023.02.015

Ranjitkar, P., Nakatsuji, T., & Asano, M. (2004). Performance evaluation of microscopic traffic flow models with test track data. *Transportation Research Record Journal of the Transportation Research Board*, *1876*(1), 90–100. https://doi.org/10.3141/1876-10

Regan, M. A., & Hallett, C. (2011, January 1). *Driver distraction: Definition, mechanisms, effects, and mitigation*. Elsevier Inc. https://doi.org/10.1016/b978-0-12-381984-0.10020-7

Saifuzzaman, M., & Zheng, Z. (2014). Incorporating human-factors in car-following models: A review of recent developments and research needs. *Transportation Research Part C Emerging Technologies*, *48*, 379–403. https://doi.org/10.1016/j.trc.2014.09.008

Saifuzzaman, M., Zheng, Z., Haque, M. M., & Washington, S. (2015). Revisiting the task–capability interface model for incorporating human factors into car-following models. *Transportation Research Part B Methodological*, *82*, 1–19. https://doi.org/10.1016/j.trb.2015.09.011

Saifuzzaman, M., Zheng, Z., Haque, M. M., & Washington, S. (2017). Understanding the mechanism of traffic hysteresis and traffic oscillations through the change in task difficulty level. *Transportation Research Part B Methodological*, *105*, 523–538. https://doi.org/10.1016/j.trb.2017.09.023

Spall, J. C. (2005, March 11). *Introduction to stochastic search and optimization: Estimation, simulation, and control*. John Wiley & Sons.

Stanton, N., Young, M., & McCaulder, B. (1997). Drive-by-wire: The case of driver workload and reclaiming control with adaptive cruise control. *Safety Science*, *27*(2-3), 149–159. https://doi.org/10.1016/s0925-7535(97)00054-4

Tang, Quinteiro, Dopico, García, & Ahlström. (2023, October 26). *I4driving causal relationships between human/external factors and human drivers behaviors: Modelling requirements and framework of testable hypotheses* (No. D1.5). https://i4driving.eu/deliverables/

Van Lint, J., & Calvert, S. (2018). A generic multi-level framework for microscopic traffic simulation—theory and an example case in modelling driver distraction. *Transportation Research Part B Methodological*, *117*, 63–86. https://doi.org/10.1016/j.trb.2018.08.009

Victor, Engström, & Harbluk. (2008, October 15). *Distraction assessment methods based on visual behavior and event detection*. CRC Press. https://books.google.nl/books?id=o7--7AS38tYC

Welch, B. L. (1947). The generalization of "student's" problem when several different population variances are involved. *Biometrika*, *34*(1-2), 28–35. https://doi.org/10.1093/biomet/34.1-2.28

Wiedemann, R. (1974). Simulation des straßenverkehrsflusses. https://trid.trb.org/view/596235

Wikipedia contributors. (2024, December 11). *Welch's t-test*. https://en.wikipedia.org/wiki/Welch's_t-test

Wikipedia contributors. (2025, March 28). *Student's t-distribution*. https://en.wikipedia.org/wiki/Student%27s_t-distribution

Young, K. L., Osborne, R., Koppel, S., Charlton, J. L., Grzebieta, R., Williamson, A., Haworth, N., Woolley, J., & Senserrick, T. (2019). What contextual and demographic factors predict drivers' decision to engage in secondary tasks? *IET Intelligent Transport Systems*, *13*(8), 1218–1223. https://doi.org/10.1049/iet-its.2018.5546

Young, K. L., Regan, M. A., & Lee, J. D. (2008, October 15). *Factors moderating the impact of distraction on driving performance and safety*. CRC Press. https://doi.org/10.1201/9781420007497-30

Unpublished (2025-06), 1–15

The Systematic Modelling of Distraction in TCI Traffic Models

F.D. Dutruel

Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, The Netherlands

Abstract

Human factors have become increasingly central in traffic modelling, such as in efforts to simulate driver behaviour under varying conditions such as cooperation and distraction. While motivations for this research are diverse, the underlying goal is consistently road safety. The Task-Capability Interface (TCI) model is a promising framework in this domain, offering a structured method to represent cognitive processes as proxies for human behaviour. Although the TCI model has been used to study distraction, prior studies have approached the topic in isolated or inconsistent ways, revealing a gap in systematically modelling distraction within TCI frameworks. This thesis addresses that gap by proposing a novel distraction framework grounded in the low-level characteristics of distractions. The framework conceptualizes distractions as a process consisting of three distinct phases: trigger, intensity, and effect. To evaluate its effectiveness, the framework was integrated into the Multi-scale model, resulting in a new Distraction model. This model was then calibrated using a genetic algorithm against two datasets containing different types of distractions: a continuous mental-visual distraction and a spontaneous auditory distraction. The performance of the Distraction model was compared to specialized baseline models tailored to each dataset. Results indicate that the Distraction model performs on par with the specialized models, with notable improvements in estimating headway during individual trajectory calibration. However, this advantage diminishes when applied as a generalized calibrated model. The findings affirm that the framework functions as intended offering a systematic and flexible approach to modelling distraction, its key strengths, even if performance gains are negligible.

Keywords: Car-following, Human factors, Task-Capacity-Interface model, Distraction

1. Introduction

Microscopic traffic simulation encompasses a wide range of model types, scopes, and implementations. A growing area within this field involves models that integrate human factors (HF), extending beyond traditional approaches that focus solely on how vehicles move according to traffic rules. HF models also aim to explain why and when drivers fail to comply with these rules (Saifuzzaman and Zheng 2014).

This shift in focus has gained momentum over the past decade, largely due to advances in automation and digitalization within transportation systems (Eltoweissy, Olariu, and Younis 2010). Since traffic continues to include human-operated and semi-automated vehicles, models must account for illogical driver behaviour and potential lapses, such as distraction.

Various theoretical frameworks have emerged for implementing HF in traffic models. Some manipulate perception by introducing minimum thresholds or transforming stimuli into intuitive relative values (Wiedemann 1974; Andersen and Sauer 2007), while others incorporate subjective risk-taking to reflect individual driver's behaviour (Hamdar et al. 2008; Hamdar, Mahmassani, and Treiber 2015). Each of these models offers advantages but tends to be narrow in focus and supported by limited literature.

This paper focuses on the Task-Capability Interface (TCI) model, a flexible control framework that models the human decision-making process by simulating cognitive effort and mental workload (Fuller 2011). Rather than describing explicit behaviour, the TCI model offers generalized principles for how drivers manage tasks in relation to their perceived capabilities. Because the thought process is a foundational human factor, it can serve as a proxy to represent more complex or higher-level behaviours in TCI models. Its flexibility has led to a variety of implementations addressing topics such as risk-taking and cooperation. However, distractions, despite being a major contributor to road incidents (Regan and Hallett 2011; Gordon 2008), remain under-explored in TCI-based models. Existing research often models only a single type of distraction per specific implementation thus limiting future comparison (Van Lint and Calvert 2018; Saifuzzaman et al. 2015; Li, Li, and Ni 2020). This paper aims to address that gap by developing a systematic framework for modelling multiple, diverse types of driver distraction within a microscopic TCI-based simulation environment.

1.1 Research Objective and Research Questions

With a gap identified in the existing literature the objective of this research becomes clear. It's the development, integration and validation of a systematic modelling framework for distractions.

To help accomplish this goal the following research question is setup: *"How can a framework for systematically defining different types of distractions be implemented and validated at a microscopic scale in a task-capability interface car-following model?"*

This research question already implies that the research will be broken up into multiple steps. The first step regards the properties of the framework. Without knowing how the framework works it will be impossible to integrate it into a car-following model. The second step covers the details of the integration process, namely which car-following model is used and how it should be modified to still retain its original identity. Lastly with the integration complete it needs to be validated, the most important question to answer during this test is if the framework is able to properly model different types of distractions. The approaches used to fulfil each of these steps are explained in the methodology section.

2. Methodology

The methodology details all the steps needed to replicate the research performed in this paper. It is divided into three subsections. The first subsection covers the details of the distraction frameworks, how and why it is defined the way it is. The following subsection explains the chosen TCI model into which the framework will be integrated. In the last section the validation methods are explained.

2.1 Distraction Framework

There have been multiple studies which try to characterize distractions, each with their own approach (Regan and Hallett 2011; Young et al. 2019; Cuentas-Hernandez et al. 2023). These studies usually come to similar conclusions but none are perfect and each approach leaves some gaps which are difficult to explain.

These characterization attempts can be divided into two distinct groups based on the level of distraction characteristic they use, either low-level or high-level characteristics. Low-level groupings, such as sensory modality or internal/external sources, provide a more granular control, while high-level groupings refer to broader activities like conversation or eating. Choosing the appropriate abstraction level for the distraction framework is critical, as each has distinct benefits and limitations.

Low-level characteristics enable the modelling of complex, multi-faceted distractions by allowing each component to influence the model independently. This modularity improves flexibility, new high-level distractions composed of existing low-level traits can be integrated without full model recalibration. However, this approach is labour-intensive, requiring multiple low-level traits to accurately simulate a single real-world distraction (Regan and Hallett 2011). Moreover, empirical validation is difficult due to the interdependence of these traits and limited data availability, as most datasets classify distractions using high-level labels (Engström et al. 2013).

Conversely, models based on high-level characteristics are easier to calibrate and align with available data. Yet, they lack adaptability, as new distractions not previously defined must be added explicitly, limiting extensibility.

Given the papers objective, to support a broad range of distraction types, a low-level approach is adopted for its flexibility and potential to model diverse and novel distraction scenarios more effectively.

With the chosen characteristic abstraction level, the core framework of this paper can be defined. The distraction framework is responsible for quantifying distractions and deciding how much of an impact a distraction has on a driver. This framework is inspired from existing literature but it is mainly this paper's interpretation of distractions (Regan and Hallett 2011; Victor, Engström, and Harbluk 2008; Young, Regan, and Lee 2008; Liang and Lee 2010).

The distraction framework can be divided into three stages, the distraction trigger, distraction intensity and distraction effect. Each of these stages represents an important part of the distraction's lifecycle. These stages are in chronological order and each stage refines the distraction from global distraction parameters to personal driver distraction parameters.

The distraction trigger is the first stage in the framework. Its purpose is to define when, where and to whom a potential distraction applies. This stage aims to provide an easy way of encoding this information and alert the relevant vehicles in the model. The encoding is formally defined with the trigger function, see equation 1.

$$DT_{i,j} = f_{DT,j}(x_i, t, i) \tag{1}$$

with $DT_{i,j}$ as boolean trigger output, x_i as driver vehicle position, t as current time. Subscript i is the driver index and subscript j is the distraction index. The exact formulation of function $f_{DT,j}$ depends heavily on the type of distraction modelled.

Next is the distraction intensity. Literature shows that there are four main factors which determine the level of reaction from a driver (Victor, Engström, and Harbluk 2008; Young, Regan, and Lee 2008; Liang and Lee 2010). These are the inherent distraction strength, the driver's task saturation, the driver's distraction sensitivity and whether the distraction is spontaneous. The purpose of the intensity stage is to transform these factors into a driver-specific intensity over time graph. This can be expressed with the following formula:

$$DI_{i,j}(t) = f_{DI,j}\left(DS_j, TS_i, DDS_i, SP_j(t)\right) \tag{2}$$

with $DI_{i,j}(t)$ as specific distraction intensity over time, DS_j as inherent distraction strength, TS_i as task saturation, DDS_i as driver distraction sensitivity and $SP_j(t)$ as binary value which denotes if the distraction is spontaneous. The subscripts i and j represent the driver and distraction indexes.

The last stage is the distraction effect, it transforms the reaction intensity into an effect on driving performance. The effects of distractions can be quite varied, multiple parameters can be affected in either a positive or negative manner, as shown in literature (Liang and Lee 2010). The exact number and manner in which driving performance categories can be altered depends on the chosen car-following model. This means that the CF model has a large impact on the exact formulation of the distraction effect. That said four common categories were identified in TCI model implementations: reaction time, task demand, driver personal preference and situational awareness. A distraction effect doesn't have to influence all of these variables, instead it might influence one or two of them depending on the nature of the distraction.

The effect of distractions on reaction time is usually quite straightforward, resulting in an increase. Secondly, distractions are a perfect example of a secondary task and all tasks have an associated task demand. The last two categories allow for a more fine-tuned effect. By modifying the driver personal preferences and situational awareness the distraction can have a more direct control over the driving performance as long as the model allows for such modifications.

2.2 *Multi-Criteria Analysis TCI model*

To test and validate the distraction framework it needs to be integrated into a TCI model. To this end a multi-criteria analysis (MCA) is performed to choose a suitable TCI model. The TCI models which are evaluated are the Multi-scale model (Van Lint and Calvert 2018; Calvert, Schakel, and Van Lint 2020), the TDCF/TDIDM (Saifuzzaman et al. 2015, 2017) and FTD-LCM models (Li, Li, and Ni 2020). The MCA is performed in three steps, first a list of criteria is defined, then the individual models are evaluated and lastly they are compared.

The criteria set contains both formal model function related criteria and informal personal preference criteria. The seven criteria used for the MCA are listed below and are not ranked in any particular order. Their reason for being included and their grading method are explained in the coming paragraph.

1. Model complexity
2. Model flexibility
3. Computational complexity
4. General human factors
5. Existing distractions
6. Author contact
7. Open source

First is model complexity. A balance is sought between richness of behaviour and interpretability. Complex models can simulate nuanced driver behaviours but are more difficult to validate and calibrate. Complexity is often assessed using conceptual diagrams; models with more steps, inputs, and feedback loops are considered more complex.

In second is model flexibility. This refers to how easily the model can be adapted, particularly to include external inputs, without requiring full recalibration or redesign. Modular or robust designs are preferred, as they allow parts of the model to be modified independently.

The computational complexity is third. Efficient models are crucial for testing and iteration. Computational demands can be evaluated either through runtime in a standardized scenario or by analysing the number of mathematical operations and feedback loops per time step. Lower computational complexity enables more extensive testing.

Fourth is human factor integration. This criterion considers the model's existing capability to include human behaviours. Ideally, models allow modular inclusion of multiple human factors to reflect realistic driving behaviour. Too few human factors result in unrealistic driver behaviour, while too many may obscure the isolated effect of the human factors like distractions.

In fifth is the existing distraction implementation. This assesses whether the model already includes distractions or supports their easy addition. A positive rating is given if the model has guidelines for implementing different distraction types and possesses the variables necessary to support new distractions, such as gaze information.

In sixth is the author contact. While more subjective, fast contact with the model's author can support troubleshooting and accelerate development by providing expert insights.

Last is the open source availability. Models with publicly available code and data are preferred, as they reduce initial development effort and facilitate validation through existing test datasets.

Together, these criteria guide the selection of a model that is capable, extensible, and practical for integrating the distraction framework.

Using these criteria the three TCI models are evaluated and compared. The results of this comparison can be seen in table 1. The model which overall performs the best is the Multi-scale model. Overall this model tends to be the most balanced, providing good results in all areas while excelling in flexibility. For this reason, the distraction framework will be integrated into the Multi-scale model.

A more detailed review of the individual criteria can be found below.

Table 1. Multi-criteria analysis results of the three reviewed TCI models.

Criteria	Multi-scale	TDCF/TDIDM	FTD-LCM
Model complexity	+	++	0
Model flexibility	++	-	0
Computational complexity	+	++	-
General human factors	+	+	++
Existing distractions	+	0	+
Author contact	++	+	+
Open source	++	+	-

Overall there are two conclusions which can be drawn from the above results. First the models do indeed accomplish what they set out to do and secondly none of the reviewed models are built to specifically handle distractions.

When looking at the results it is noticeable that each of the models excels in a specific area. The Multi-scale model, which is built to be modular and serve as a base for other models, has excellent flexibility while providing good results in other areas of interests. The task difficulty model, which aims to expand upon the most commonly used and simple models, has an extremely fast computational speed and a lower level of complexity compared to the others. Lastly the Fuzzy task difficulty model, which aims to improve the human factors by adding more human-friendly inputs and fuzziness in calculations, has the best human factors.

That said while all models have a method of supporting secondary driving task and distractions none of the models focus on that aspect of driving. None of the models give guidelines on how these should be modelled or when these should occur.

Table 2. Summary of distraction and social characteristics of used datasets.

	UNINA	CARRS-Q
Trajectory duration	∼120s	20-30s
Distraction modality	Visual & mental	Auditory
Distraction duration	Full simulation	Half simulation
Spontaneous distraction	No	Yes
Participant age	19 <31 y/o & 13 >31 y/o	18-26 y/o
Participant gender	16M & 16F	equal
Participant experience	15 inexperienced & 17 experienced	mostly inexperienced
Participant cuture	Italian	Australian

2.3 Validation

The validation of the distraction framework is done by integrating it into a car-following model and testing this model. This chapter outlines the methodology used to assess the effectiveness of a newly developed Distraction model. Given the challenge of validating the model across all distraction types, the approach focuses on demonstrating limitations in existing models under specific conditions that the new model addresses more effectively. The chapter is structured in three parts: the design of the validation test, the technical integration of the distraction framework into the car-following model and the calibration tools.

The validation scenario is based on a proof-of-concept approach. Rather than a full validation for all distractions, the model is tested on two distinct types of distraction. To this end two datasets with distractions are used (dataset A and B), and evaluated with two corresponding specialized TCI models (baseline model A and B). The hypothesis is that the Distraction model should outperform or at least match each baseline model on its corresponding dataset, demonstrating broader applicability. To this end the following hypothesis are defined: the null hypothesis is that the model performs worse than the baselines and the alternative hypothesis is that it performs equally or better.

Since the model is built upon low-level characteristics, as mentioned in the distraction framework section, it means the two different distraction have different low-level characteristics. Normally the two different distractions are chosen first and corresponding datasets are found afterwards. This is not the case for this paper, instead the opposite approach is taken since data availability is the limiting factor. This means that two datasets with distractions are chosen and analysed to see if their distractions are different enough. If that is found to be the case they are accepted for the validation test.

The selected datasets are the i4Driving UNINA motorway data and CARRS-Q data. The distractions in these two datasets are a continuous mental-visual distraction and a spontaneous auditory distraction, a phone call. The main low-level attributes of the mental distraction are as follows: an internal distraction, a voluntary distraction and a visual-mental sensory modality. On the other hand the low-level attributes of the phone call are: internal distraction, an involuntary distraction and auditory sensory modality. These two distractions have different low levels characteristics and are thus suitable for the validation test. Both datasets contain multiple scenarios in which drivers perform the tests under varying conditions. For the UNINA dataset scenario 4 and 5 are used, these two scenarios are identical experiments based on real-world motorway data in a single lane straight uninterrupted segment. The difference between the two scenarios is the presence of the distraction; in scenario 4 no distraction is present and in scenario 5 the distraction is present from start till end. The CARRS-Q dataset follows a similar setup to the UNINA data. The main difference is the lower speed due to being in a urban setting. The used scenarios are scenario 1 and 3. Scenario 1 contains no distraction while scenario 3 contains a handheld phone call half way through the simulation. More details regarding the characteristics and social factors of these datasets are compiled in table 2.

The Multi-scale framework is employed to represent the continuous distraction observed in the i4Driving UNINA dataset, as it is specifically designed for mental distraction scenarios. In contrast, the TDIDM model is used for the CARRS-Q dataset, which features spontaneous auditory distractions, due to its prior training and effective performance on this data. While the TDIDM model requires only calibration, the Multi-scale model necessitates additional configuration for the distraction task demand curve. A theoretical reference model, termed the Multi-scale Five model, is also introduced as benchmark for task demand estimation. Unlike the baseline models, this reference model disregards the scenario conditions and distributes distraction-related task demand across five equal time segments, enabling refined analysis.

In order to determine if the Distraction model performs better than the other models the resulting trajectory of the model is evaluated. To measure how good a trajectory is we compare it to the real measured trajectory from the original data. The closer the simulated trajectory is to the measured trajectory the better. To do this the symmetric mean absolute percentage error (SMAPE) of the spacing, speed and time headway is calculated for each simulated trajectory, see equation 3. Ideally, to conclude that model A is better than model B, the three SMAPE's of model A should be statistically lower than model B. That said the minimum needed to conclude that model A is better than model B would be for one of the

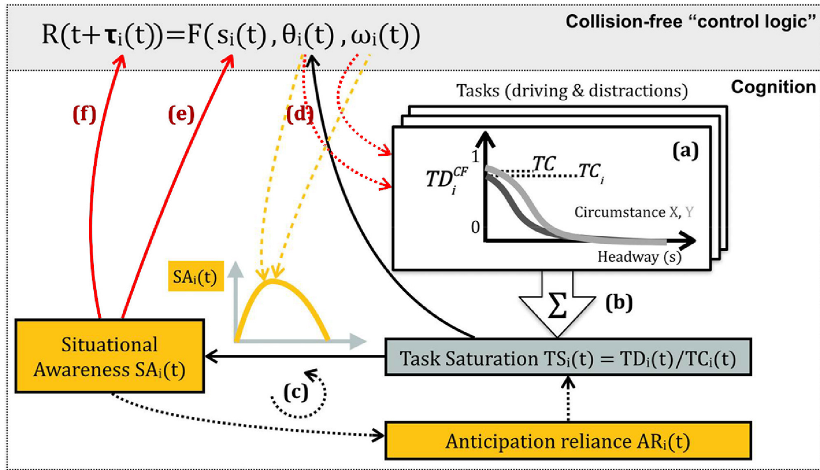

Figure 1. Multi-scale model relationships modified by the distraction framework in red. Figure source: (Calvert, Schakel, and Van Lint 2020) (Modified)

SMAPE's to be statistically lower than the other while the other two are statistically equivalent. In the case that at least one of the SMAPE's is statistically lower and at least one is statistically higher then the results are inconclusive and no model can be declared as superior. These statistical differences are determined with a one-tailed Welch's t-test

$$SMAPE_h = \frac{1}{N}\sum_{n=1}^{N}\frac{||h_{sim} - h_{real}||}{||h_{sim}|| + ||h_{real}||} \tag{3}$$

To validate the distraction framework it needs to be fully integrated into the Multi-scale model. The following paragraphs will explain how this is done and what equations are used for the integrations. To start an overview of the modified section of the Multi-scale model are given in figure 1. The arrows in red are the sections of the Multi-scale model which are modified. Arrow (d) is the connection between driver preference ($\theta_i(t)$), environment ($\omega_i(t)$) and task demand calculations. Arrows (e) and (f) are the connections between situational awareness ($SA_i(t)$) and reaction time ($\tau_i(t)$) and perceived stimuli ($s_i(t)$).

The first two stages, distraction trigger and intensity, take place in arrows (d) in figure 1. The trigger depends heavily on the type of distraction, so the two datasets get their own trigger implementation, see equations 4 and 5.

$$DT_i(x_i, t) = \left(\sum_{t_g=t-t_{linger}}^{t} gaze_{dis}(t_g)\right) > 0 \tag{4}$$

$$DT_i(x_i, t) = t > t_{start,\ i} \tag{5}$$

with t_{linger} as distraction linger time, $gaze_{dis}(t)$ gaze direction distraction and t_{start} as distraction start time.

The motivations behind the equations are as follows. The UNINA visual distraction is only applicable when the driver looks at the screen with the question and a short time afterwards while the driver answers the question. The exact manner of incorporating gaze data into the formula is not based on literature since no relevant literature was found using gaze data in car-following models or other operational-level decision making models. The only model which applied gaze data are tactical-level decision making models. The CARRS-Q distraction is dependant on a spontaneous phone call, this is far simpler to model since only the start time of the first phone ring is needed and is provided by the dataset.

These two equations also make use of a couple of constants, namely the linger time and the start time. The linger time is set to 2 seconds, multiple options were tried and 2 seconds provides the best and most consistent results. The start time is simulation/driver specific, it is provided by the CARSS-Q dataset and is setup during initialisation of the scenario-driver pair.

The second stage, distraction intensity, is implemented in equation 6.

$$DI_i = \begin{cases} TD_{d,\ i} = DS * DDS_i & SP = 0 \ \vee t > t_{start,\ i} + 2s \\ TD_{dsp,\ i} = DS * DDS_{SP,\ i} & SP = 1 \ \wedge t > t_{start,\ i} \end{cases} \tag{6}$$

$$TD_{total,\ i} = DI_i + TD_{cf} \tag{7}$$

with $TD_{d,\ i}$ as distraction task demand, $TD_{dsp,\ i}$ for spontaneous distraction task demand, free parameter $DDS_{SP,\ i}$ as spontaneous driver distraction sensitivity and $t_{start,\ i}$ as distraction start time.

The logic behind this equation depends mostly on whether a distraction is spontaneous or not. In case of a normal distraction the distraction intensity is a standard constant rate. If the distraction is spontaneous it gets a modified rate for two second and afterwards it returns to normal distraction rate. The two seconds duration isn't chosen at random but is instead based on the average time it takes to pickup the phone call found in the data. Furthermore since the distraction strength is a constant and the driver only experiences one type of distraction during the simulation it can be simplified into a singular value which for ease of use is called distraction task demand. The second equation, 7, shows how the distraction intensity is

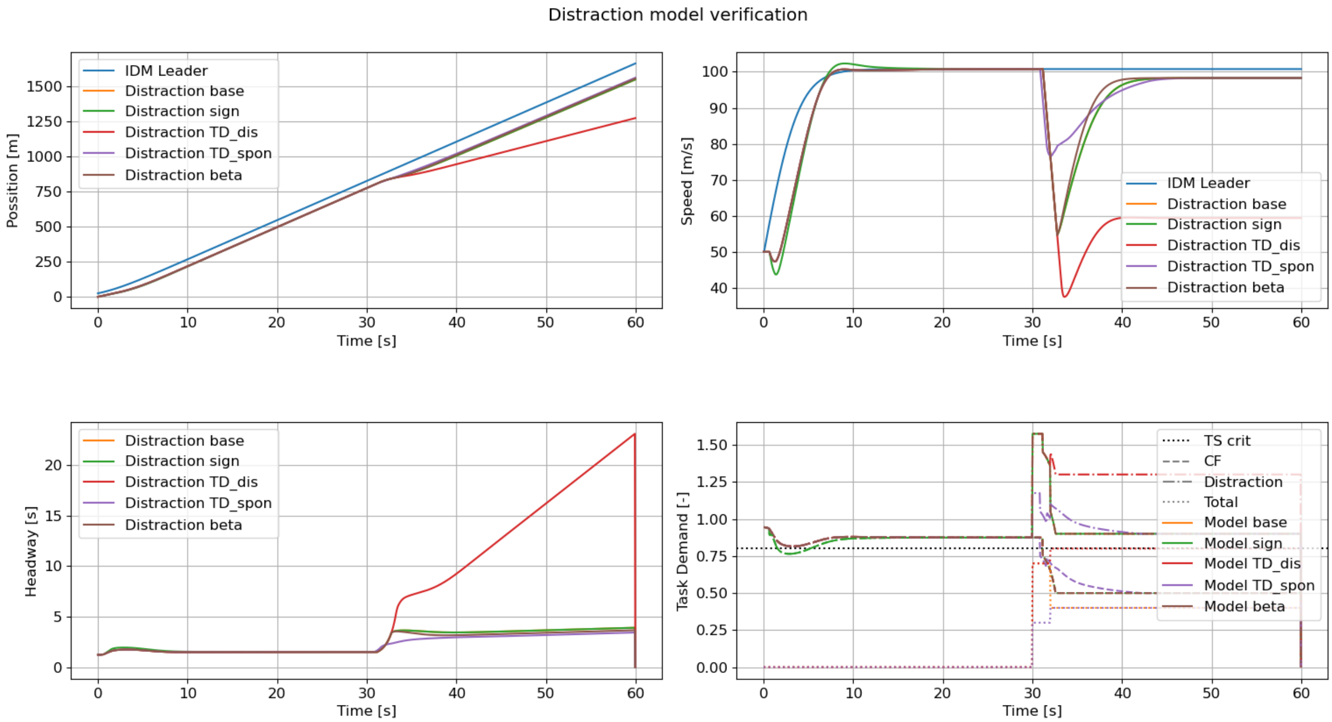

Figure 2. Trajectories of Distraction model and verification variants.

implemented in the Multi-scale model. It is simply added to the other task demands to calculate the total task demand.

The distraction intensity equation makes use of a couple of constants. These are the distraction strength, which is set to 1 for convenience since only a single distraction is present in the simulations. The two driver distraction sensitivity parameters which are estimated during calibration and are driver specific. The spontaneous factor (*SP*) which represents if a distraction is spontaneous or not and is tied to the distraction/scenario. Lastly the distraction start time is used in the same manner as in equation 5. For UNINA it is simply 0 since the distraction starts right away and for the CARRS-Q dataset it is the start of the phone call.

The last equation is a form of distraction effect. In the standard formulation of the Multi-scale model there are already a few equations which could be considered distraction effects, namely the driver preference and reaction time modifications. The Distraction model adds onto those equations with equation 8. This equation is located at arrow (e) in the logical flow diagram in figure 1.

$$a_{max,\,i}(t) = (1 - \beta_{a,\,i} * DI) * a_{max} \tag{8}$$

with free parameter β_a as maximum acceleration reduction factor and $a_{max,\,i}(t)$ as specific maximum acceleration.

This equation was added with some trial and error after analysing the predicted trajectories and comparing them to the real trajectories. One notable reoccurring flaw was that the vehicle accelerations were too high when under a high task demand. This effect is further supported by the correlation between the effort and the absolute acceleration. This correlation has a R-squared of 0.24, while weak it does show that there is a connection between the two. Multiple variants were tried each with their own pros and cons, the best performing variant is the version which is listed above.

This equation uses a single new constant, the maximum acceleration reduction factor. It is a driver specific parameter that is determined during the calibration process.

Before the Distraction model is used in the validation test it needs to be verified. This means that each new individual parameter behaves and modifies the model as expected. This is done by modifying one parameter at a time and comparing the new output to an unmodified version of the Distraction model. The parameter values and altered values can be seen in table 3. The output trajectories of the verification test can be seen in figure 2. It shows that each parameter behaves as expected and that the Distraction model passes the verification test.

To accurately replicate real trajectories, the used models must be calibrated using optimization techniques. This paper employs a Genetic Algorithm (GA) for parameter calibration due to its robustness in avoiding local minima through stochastic global search methods (Kesting and Treiber 2008; Punzo, Ciuffo, and Montanino 2012). While computationally intensive, this drawback is mitigated via code optimizations and multi-threading.

Table 3. Distraction model parameters for verification.

Parameter	Value	Altered value	Unit
b_{max}	8	-	m/s^2
a_{max}	4	-	m/s^2
b_{comf}	3	-	m/s^2
v_0	30	-	m/s
s_0	8	-	m
T	1.2	-	s
τ	0.5	-	s
$\delta\ sign$	0	1	–
TD_d	0.4	0.8	–
TD_{dsp}	0.7	0.3	–
β_a	0.6	0.2	–

A critical aspect of GA calibration is the objective (fitness) function, which evaluates model performance. Though multiple metrics (SMAPE of position, speed, and time headway) are used in validation, only a single metric can serve as the objective. The SMAPE of headway was selected, as it implicitly captures both position and speed errors, offering the most comprehensive performance measure (Kesting and Treiber 2008).

Additional GA parameters include population and gene settings. Population parameters, number of generations (250), parents mating (4) and population size (8), determine the search scope. An early-stopping criterion halts optimization if no improvement occurs over 50 generations. Genes represent the model parameters to be optimized. Two key gene parameters are the mutation rate (25%) and gene bounds, the latter defining the feasible search space. Gene bounds are model-specific and detailed separately for each model, they can be seen in tables 4, 5, 6, they originate from a study by Papathanasopoulou and Antoniou (2015), physical limitations and theoretical limitations. The *pygad* Python package is used to implement the GA.

Table 4. Common model free parameter bounds for GA optimisation. *Fixed values, based on test data.

Parameter	Min	Max	Step	Unit
b_{max}*	8	8	-	m/s^2
a_{max}*	4	4	-	m/s^2
b_{comf}	0.5	5	-	m/s^2
v_0	10	50	-	m/s
s_0	1	15	-	m
T	0.5	5	-	s
τ	0.1	2	-	s

Table 5. TDIDM model free parameter bounds for GA optimisation (Saifuzzaman et al. 2015).

Parameter	Min	Max	Step	Unit
δ	-5	1	-	–
γ	0.5	5	-	-
φ	0.1	2	-	s

Table 6. Multi-scale and Distraction model free parameter bounds for GA optimisation.

Parameter	Min	Max	Step	Unit
$\delta\ sign$	0	1	1	-
TD_d	0	1	-	-
TD_{dsp}	0	1	-	-
β_a	0	1	-	-

Before using the designed GA in the validation test it needs to be verified with an existing validated reference. To this end a synthetic data test is performed. This means that the GA is used to estimate the parameters of a trajectory made by a car-following model. The model used is the Multi-scale model. The model parameters and test results can be seen in table 7. These results show that the GA performs well and that it is able to accurately estimate nearly all parameters.

Table 7. Synthetic Multi-scale model optimisation test parameters. *Fixed parameters. **Not present in scenario.

Parameter	Synthetic	Scenario distraction	Scenario slowdown	Scenario complex
b_{max}* $[m/s^2]$	8	-	-	-
a_{max}* $[m/s^2$	4	-	-	-
b_{comf} $[m/s^2]$	3	2.94 (-2.0%)	3.58 (+19.3%)	2.99 (-0.3%)
v_0 $[m/s]$	35	34.99 (-0.0%)	35.06 (+0.2%)	34.99 (-0.0%)
s_0 $[m]$	8	8.25 (+3.1%)	1.71 (-78.6%)	11.17 (+39.6%)
T $[s]$	1.2	1.10 (-8.3%)	1.48 (+23.3%)	1.06 (-11.7%)
τ $[s]$	0.5	0.51 (+2.0%)	0.48 (-4.0%)	0.50 (+0.0%)
$\delta\ sign$ [-]	0	0.5 (0.0%)	0 (0.0%)	0 (0.0%)
TD_d [-]	0.7	0.70 (0.0%)	-**	0.70 (0.0%)

3. Results

Using the methods described in the previous section the data and Distraction model are evaluated. This means that both the individual trajectory calibration results and the calibrated model results are presented in this section. That said the majority of the section will focus on the calibrated model results.

3.1 Individual Trajectory Results

A brief overview of the individual trajectory calibration results is given in the following paragraphs.

Table 8. Calibrated model average KPI, all scenario's all seeds.

Scenario & KPI	Models & Seeds											
	IDM			TDIDM			Multi-scale			Distract model		
	1	2	3	1	2	3	1	2	3	1	2	3
UNINA 4												
$SMAPE_x$ [m]	96.6%	97.0%	91.3%	86.8%	93.3%	88.0%	90.2%	96.9%	91.3%			
$SMAPE_v$ [m/s]	95.1%	96.0%	93.5%	94.1%	95.2%	92.9%	94.6%	95.7%	93.3%			
$SMAPE_h$ [s]	75.4%	78.6%	70.6%	61.8%	71.2%	62.0%	71.2%	79.2%	71.2%			
UNINA 5												
$SMAPE_x$ [m]	98.0%	98.2%	98.0%	97.9%	98.2%	97.9%	98.1%	98.3%	98.0%	97.6%	98.0%	98.0%
$SMAPE_v$ [m/s]	91.5%	91.4%	91.5%	91.5%	91.4%	91.5%	91.7%	91.5%	91.3%	90.1%	90.4%	90.8%
$SMAPE_h$ [s]	78.9%	80.5%	79.3%	78.8%	79.0%	78.6%	78.8%	80.5%	78.9%	77.3%	79.6%	78.8%
CARRS-Q 1												
$SMAPE_x$ [m]	95.5%	97.1%	96.7%	94.4%	95.9%	95.3%	96.2%	97.3%	97.1%			
$SMAPE_v$ [m/s]	94.8%	96.0%	95.6%	90.1%	92.9%	89.2%	94.5%	95.3%	95.2%			
$SMAPE_h$ [s]	75.7%	81.8%	83.6%	73.6%	78.2%	78.0%	81.2%	82.6%	84.7%			
CARRS-Q 3												
$SMAPE_x$ [m]	95.9%	96.9%	96.1%	93.5%	95.8%	94.4%	95.8%	96.9%	96.9%	95.7%	96.8%	96.5%
$SMAPE_v$ [m/s]	94.9%	95.0%	94.7%	90.2%	92.6%	90.8%	93.6%	94.6%	94.4%	93.1%	93.8%	93.7%
$SMAPE_h$ [s]	80.1%	81.7%	80.7%	69.9%	79.6%	73.7%	78.8%	80.9%	84.0%	76.6%	80.5%	79.6%

The first set of evaluated trajectories are the UNINA scenario 4 trajectories. The results of these calibrations show that all models perform similarly for all three KPIs. The only differences between the models are their distribution of results or standard deviation with the TDIDM having a slightly tighter distribution. While the models perform well they haven't reached the theoretical maximum estimated by the Multi-scale Five model which scores a few percentage higher for all KPIs. This difference is also large enough to make the Multi-scale Five model significantly better than the baseline models.

A similar pattern is also shown in UNINA scenario 5. All used models perform worse than in scenario 4 for nearly all KPIs, this is expected since a distracted trajectory is harder to estimate. Unlike scenario 4 this scenario shows that the IDM model seems to perform slightly better in terms of headway than the other baseline models. This increase in performance is matched by the Distraction model. Despite the difference in performance this gap isn't large enough to call the increase statistically significant. Finally the best performing model is still the Multi-scale Five model. Compared to scenario 4 the gap between it and the other models has decreased a bit but it is still considered significantly better for most KPIs (except speed).

The analysis of the results of CARRS-Q scenario 1 are slightly different from the results in UNINA. First the CARRS-Q calibration performed better overall compared to the UNINA calibration, this increase is especially pronounced for the headway. Unlike the UNINA calibration, one model stands out as the best model, namely the IDM model. This is surprising since the TDIDM model was built for this dataset. This increase in performance from the IDM model is statistically significant and nearly matches the performance of the Multi-scale Five model.

The last set of trajectories is from CARRS-Q scenario 3. Like scenario 1 the best model is still the IDM but it is no longer significantly better. This is likely due to the loss in performance caused by the distraction which affected the IDM model the most. This also means that the gap between the baseline models and the Multi-scale Five model increased in scenario 3 compared to scenario 1. Another result which increased for all the models is the standard deviation. Due to this increase the Multi-scale Five model is no longer significantly better than the baseline models. The last model to consider in scenario 3 is the Distraction model. Its performance is quite good. It manages to match the Multi-scale Five model for the headway KPI and otherwise matches the Multi-scale model for the speed and position KPI.

3.2 Calibrated Model Results

The following subsection covers the results of the calibrated models. This includes the calibrated model parameters, average KPI results, the statistical t-tests and a more in-depth look at some of the individual results.

For the UNINA dataset the parameter differences between scenarios 4 and 5 align with expectations. The desired speed remains consistent across scenarios, except for slightly lower values in Multi-scale models. Spacing at standstill generally decreases under distraction, likely due to delayed driver reactions, except in the TDIDM model. This reduction is offset by an increase in desired headway, $\sim$1 second, consistent with task difficulty homeostasis theory (Fuller, McHugh, and Pender 2007). HF models, TDIDM and Multi-scale, predict larger headways than IDM, reflecting their human factor integrations. Reaction times are mostly stable, with a slight decrease

Table 9. Calibrated model t-test, Distraction model vs other models - all seeds. Passed the 5% probability threshold.

Scenario & KPI	Distraction model vs								
	IDM			TDIDM			Multi-scale		
	1	2	3	1	2	3	1	2	3
UNINA 5									
$SMAPE_x$ [m]	0.871	0.734	0.550	0.811	0.737	0.448	0.925	0.837	0.610
$SMAPE_v$ [m/s]	0.873	0.832	0.761	0.877	0.813	0.773	0.912	0.847	0.688
$SMAPE_h$ [s]	0.754	0.688	0.575	0.741	0.402	0.469	0.748	0.684	0.511
CARRS-Q 3									
$SMAPE_x$ [m]	0.590	0.552	0.274	**0.016***	**0.033***	**0.002***	0.577	0.582	0.761
$SMAPE_v$ [m/s]	0.951	0.999	0.922	**0.004***	**0.005***	**0.001***	0.672	0.973	0.846
$SMAPE_h$ [s]	0.833	0.620	0.612	**0.027***	0.397	**0.049***	0.760	0.539	0.863

in scenario 5 for the Multi-scale model. The distraction task difficulty parameter is higher in the Distraction model than in the Multi-scale model, likely due to its intermittent presence.

Inter-seed variability in UNINA is minimal, suggesting robust driver behaviour representation, with only isolated anomalies, for example spacing or preferred braking values in specific seeds.

As for the CARRS-Q dataset greater variability is observed across models and seeds compared to UNINA. Desired speed varies significantly, especially in the Multi-scale model which has a desired speed of 33 m/s vs. 13 m/s in IDM, potentially due to underlying driver preference reduction in the Multi-scale model. Spacing at standstill remains consistent across scenarios and reflects the same hierarchy found in UNINA (TDIDM > Multi-scale > IDM). Desired headways again increase under distraction and are higher in HF models. Reaction times are slightly lower overall, $\sim$0.4 s vs. 0.5 s in UNINA, with an unexplained anomaly of $\sim$0.2 s in the Multi-scale model under distraction. The distraction task demand parameter in the Distraction model shows higher values for continuous distraction than for spontaneous distractions. This could be due to some drivers who show little reaction to spontaneous stimuli such as a ringing phone, resulting in lower average values than expected.

Concerning dataset variability, UNINA shows low sensitivity to seed selection, 5–10% parameter variation, while CARRS-Q shows higher variability, 20–30%, suggesting more heterogeneous driver behaviour likely due to younger, less experienced participants in the CARRS-Q dataset.

Table 8 shows the average KPIs of the validation data split for the calibrated models. This is one combined overview which shows all models, all scenarios and all seeds. A more in-depth example look at the results can be seen in figure 3. When compared with the KPIs of the individual trajectory calibrations, the average values of the calibrated models are much lower. This is to be expected since generalised calibrated models score worse on known data but better on unknown data than individual trajectory models. On average the speed and spacing KPIs of all models lost about 5% compared to individual trajectory KPIs. This loss in performance is even greater for the headway KPI, most models lost around 10% and the Distraction model lost about 15% in all scenarios. This additional loss in performance by the Distraction model doesn't seem to be arbitrary since the headway was also the best performing KPI of the Distraction model compared to the baseline models. It is likely that the headway estimation of the Distraction model was overfit for the individual trajectories and this over-fitting was lost when the model parameters were averaged.

In the end it would seem that all models are roughly equal. For some scenarios some of the models score 1 or 2% better than the Distraction model but this depends on the seed so it is difficult to draw a strong conclusion. Most models aren't too sensitive to the changes in seed, and subsequently different data splits, but the exception to this seems to be the headway of the TDIDM model. Where most models have a KPI difference of at most 5% due to changes in seed for the TDIDM headway this difference is between 5 and 10% for every scenario.

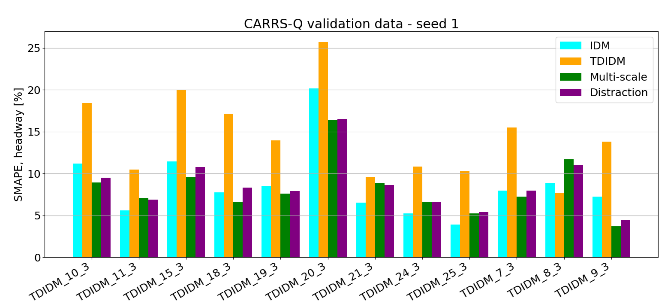

Figure 3. CARRS-Q Scenario 3 Seed 2 validation datasplit, calibrated model headway results.

Using the KPI results of the validation data split a set of t-tests are performed. These t-tests compare the Distraction model against the other models for both scenarios with distraction for all seeds. The results of these t-tests can be seen in table 9.

The results show that all the models are statistically equivalent for the UNINA dataset. This is the same conclusion as was found for the individual trajectories. So the Distraction model is not better at predicting driver behaviour than the baseline models when the driver is experiencing continuous mental/visual distractions.

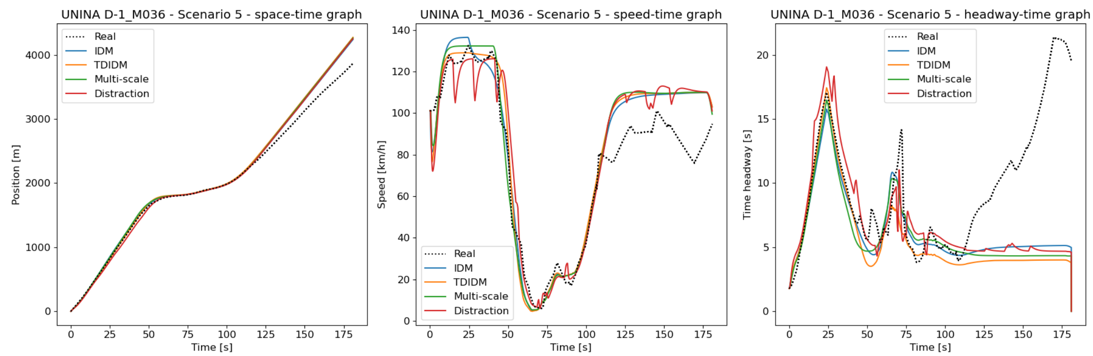

Figure 4. Trajectories of all models, UNINA validation example.

The CARRS-Q dataset does offer a slightly more diverse set of results. The t-tests show that the Distraction model is not always equivalent to the baseline models. To start the Distraction model is almost always significantly better than the TDIDM with the exception of the headway KPI in seed 2. According to the assessment rules this means that the Distraction model is significantly better than the TDIDM model for all three seeds. When comparing the Distraction model to the IDM and Multi-scale models the t-test shows that the results of the spacing and headway KPI are equivalent. As for the speed KPI it shows that in most cases the Distraction model is significantly worse. Since only one of the KPI is significantly worse and the other two are equivalent it can't be said that the Distraction model is significantly worse than the IDM or Multi-scale models. Overall these results differ quite a bit from the individual trajectory results. In the individual trajectory results the Distraction model performed significantly better on headway than the other models but this performance was lost, instead it gained a significantly better performance on speed and spacing over the TDIDM model.

3.3 Simulated Distraction Effect Analysis

The global model performance analysis given in the previous section gives us a good idea of which model performed best but it doesn't explain why this is the case. To this end a more detailed analysis is needed. To do this some of the individual validation trajectories of the calibrated models are analysed in more detail.

The trajectories which were analysed are from data-split distracted scenarios seed 1. Representative trajectories of each dataset are presented in figure 4, 5 and 6.

The first model which is analysed is the IDM model. On average this model performed the best in terms of KPI and an explanation for this behaviour can be seen in the above graphs. The IDM is overall the most stable model, this is both a positive and a negative aspect. This stability means that there are few fluctuations, big or small, in the driving behaviour of the model. This results in really straight and consistent speed and headway curves. A good example of this can be seen in figure 5 and 6. These stable curves result in good scores for the KPI calculations since its never too far of the real trajectory. That said stable curves are contrary to the objective of HF models. The goal of HF factor models is to explain the dynamic fluctuations in driving behaviour seen in real trajectories. Naturally this is difficulty to achieve with a calibrated model since a calibrated model model can only capture general trends and not individual specific trends.

Another positive aspect of the IDM model is that it has the smallest overshoots. When strong accelerations or decelerations periods occur all models tend to overshoot the performance of the real trajectory, IDM model included. In reality the driver seems to anticipate the end of the acceleration periods and reacts pre-emptively or near instantaneously. Since the models aren't built with this type of anticipation in mind it results in overshoots. That said the IDM model reacts and recovers the fastest. This is likely due to the lack of reaction time in the IDM model leading to instantaneous changes in accelerations at the end of the acceleration periods. An example of this can be seen in the speed graph of figure 6 at 25 seconds.

The model which performed the worst in terms of KPI is the TDIDM model. The reason for this is quite clear when analysing the individual trajectories. Simply put the TDIDM model is too sensitive and unstable. The performance of the TDIDM model can be separated into three cases. In the first case the model performs well without any instability, see figure 4. Its performance is similar to the IDM model and in the example it is even slightly better. That said these stable cases are quite rare; in the second and majority of the cases the TDIDM model reacts as seen in figure 5. In these cases a large spike in speed at the start of the simulation causes the model to be offset for the rest of the simulation. The model tends to recover relatively slowly which causes a large impact on the KPI result of the model. In the third most extreme and least common case the model is completely unstable, see figure 6. The model simply never recovers from its initial mistakes and sometimes even amplifies the issue resulting in a terrible KPI. The reason why this model is so unstable is not entirely clear. It doesn't seem to be related to the calibration process as multiple methods of calibration were tested yielding no improvements; the issue seems to be linked to the inherent nature of the model.

In terms of HF the model is slightly better than the IDM model. The model is able to react to some of the smaller fluctuations in

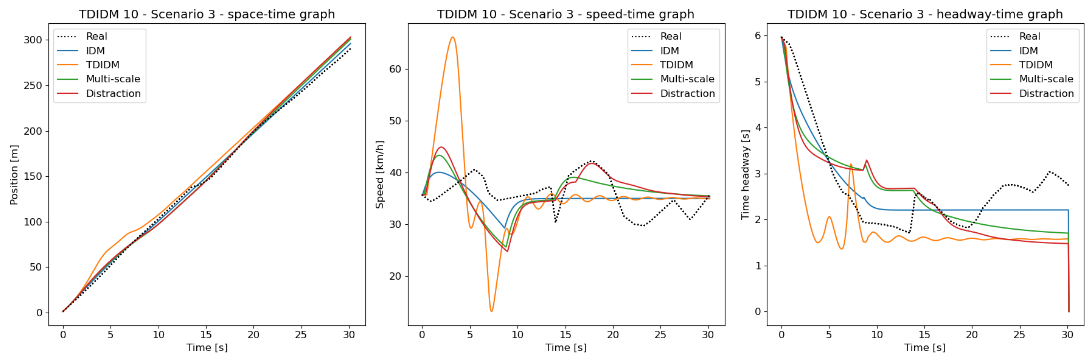

Figure 5. Trajectories of all models, CARRS-Q validation example 1.

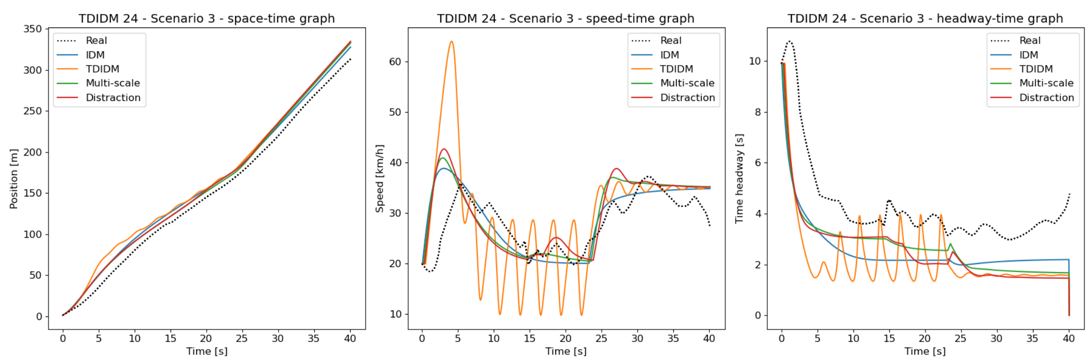

Figure 6. Trajectories of all models, CARRS-Q validation example 2.

driving behaviour but it doesn't react to every fluctuation and when it does it greatly overestimates the response, see figure 5 speed curve around 6 seconds as example.

Compared to the previous two models the Multi-sale model behaves much closer to what would be expected of a HF model. In both the UNINA and CARRS-Q datasets it is able to predict some of the smaller fluctuations in driving behaviour. In the UNINA trajectories this is mainly visible in the headway curves. It seems to follow the trends in trajectory better than the IDM model though the differences aren't large. The CARRS-Q trajectories is where the Multi-scale model shows a major difference compared to the IDM model. Since the Multi-scale model knows when the phone calls start it is able to change its behaviour at that moment and this change in behaviour results in a notable jump in the speed and headway curves. These reactions are quite similar to the reactions seen in the real trajectory and are not captured at all in the IDM model. While these reactions happen at the correct moment and are of the correct type they have the wrong amplitude. As a result the trajectories of the real car and simulated car become slightly offset resulting in worse KPI for the rest of the simulation. This leads to the second big difference when compared to the performance of the IDM model. The general performance of the Multi-scale model is a bit worse than the IDM model; even without taking the offsets into account it seems that the Multi-scale model is a bit more conservative in its headway estimates. This can be beneficial as seen in 6 or detrimental as seen in 5 and on average this seems to be detrimental as reflected by the performance KPI. So in conclusion, the Multi-scale model is better at estimating the small fluctuations in driving performance than the previous two models but at the cost of general accuracy.

The last model is the Distraction model. With its more specific distraction triggers and distraction effects it is expected that the model is able to accurately capture most of the small fluctuations in driving behaviour while also keeping a good performance of the overall picture.

Looking at the trajectories of the UNINA dataset this expectation seems to be met. An example of this can be seen in the speed graph of figure 4. It shows that the model is able to estimate the spikes in speed in the first third of the trajectory, even if their amplitudes are on the low side. A similar performance can be seen in the last third of the trajectory where it is able to predict changes in behaviour at the right moment but the amplitude and type of reaction are incorrectly estimated. The middle third of the trajectory shows that the Distraction model has more difficulty estimating the low speed distraction effect. Instead of resulting in a decrease in speed it instead results in an increase in speed. These types of predicted effects shown in the example happen in most of the trajectories of the validation set.

Similar behaviour is also seen in the CARRS-Q dataset. The Distraction model behaves quite similarly to the Multi-scale model but its two different distraction levels allows it to more accurately estimate the changes in speed that occur at and after the phone call. This can be seen in figure 5 around the second 20 mark and in figure 6 at the 18 second mark. That said this increase in performance again comes at the price of general performance.

4. Discussion

The discussion covers all areas of improvement identified in the research which haven't been studied in detail. It briefly covers what they are, what impact they could have on the conclusion and how they should be mitigated in the future.

4.1 Methodology

The methodology section of this thesis involves several critical decisions, primarily grounded in literature and empirical testing.

A central decision concerns the structure of the distraction lifecycle, which is divided into three stages to balance clarity and flexibility. While fewer stages would simplify the framework, and more would increase granularity, three stages provide a practical compromise. For instance, separating distraction intensity and effect allows the model to distinguish between distraction meta-parameters and distraction driver behaviour, enhancing modelling capabilities despite some conceptual overlap. The chosen structure remains open to future extensions should novel distraction mechanisms emerge.

Another key decision was the selection of an existing car-following (CF) model to integrate the distraction framework. A custom CF model could offer better integration of distraction data, like gaze data, and potentially improve realism but would require extensive development with uncertain gains. Car-following model evaluations research by Saifuzzaman and Zheng (2014) suggests limited improvements from newer CF models, therefore the thesis opts for a proven model with minimal modifications to ensure robustness and avoid unnecessary revalidation.

The validation methodology was also constrained by data availability. While the research question encompasses all distraction types, only two datasets were used. A broader validation approach, if more data were available, would involve assessing the framework's ability to represent different distractions, thereby creating reusable lifecycle components. This modular concept remains a promising avenue for future work.

The implementation of the distraction model involved selecting among several variants for each lifecycle stage. Choices were informed by literature and practical testing, though only the most relevant variants for the limited distraction types were retained. This highlights the need for future research into which lifecycle components are generalizable across distraction types.

Finally, a genetic algorithm was chosen for model calibration due to its widespread use in traffic modelling literature and suitability for complex optimization tasks (Kesting and Treiber 2008; Spall 2005; Ranjitkar, Nakatsuji, and Asano 2004). Though alternative stochastic algorithms like simulated annealing and particle swarm optimization exist, the genetic algorithm's documented effectiveness and available optimization guidelines made it the most appropriate choice for this study.

4.2 Data

Regarding the UNINA dataset, two major issues were identified. First, abnormally high headways were observed, averaging 6 to 10 seconds in unfiltered trajectories and $\sim$5 seconds in car-following sections, well above the typical car-following headway of 0.5 to 2.5 seconds (Saifuzzaman and Zheng 2014). This led to inflated model parameters, for example, headways of 3 to 4 seconds and standstill distances of 10 to 20 meters. While not affecting internal comparisons, these unusual deviations limit comparability with other studies. The high headways also likely reduce relative error metrics, skewing performance evaluations. Second, the assumed correlation between task difficulty and physiological effort provided by the model, heart rate, EEG and pupil diameter, appears inconsistent. Despite scenario 5 involving distractions, effort levels remained relatively flat, and no measured effort peaks aligned with known distraction events. This raises doubt about the robustness of the effort–task difficulty correlation proposed by i4Driving (Tang et al. 2023; Irvine et al. 2023) or suggests that the distractions used lacked sufficient cognitive impact. Further research with stronger or mandatory distractions is recommended.

In the CARRS-Q dataset, three design-related flaws were noted. First, short trajectory durations of 20 to 30 second limited the model calibration quality compared to longer UNINA trajectories. Secondly, participant homogeneity, mostly young drivers, introduces a bias toward inexperienced driving behaviour, affecting parameter generalisability. While this does not invalidate the model comparison within the dataset, it limits broader applicability. Lastly, the instantaneous leader speed increase during simulation is unrealistic and may favour some of the used models. This could unfairly influence performance rankings. It is difficult to identify which model would benefit from such an unrealistic effect. To this end, a controlled experiment isolating this speed increase is suggested. Ideally, the experiment design should be revised to eliminate all three issues.

4.3 Results

This section discusses unexpected outcomes from the individual trajectory calibrations.

While most deviations are explainable, two key anomalies in individual trajectory calibration merit further analysis. First, the Distraction model exhibits greater performance loss in the headway KPI, roughly 15%, than other models, $\sim$10%, for the calibrated model step. This is likely due to over-fitting on individual trajectories, with potential causes including the model's complex parameter set and extensive calibration process. Unlike simpler models, the Distraction model continued marginal improvements until reaching the generation cap, possibly exacerbating over-fitting. Moreover, its five human factor parameters, more than any other model, may reduce generalisability during aggregation. However this model performance loss correlation with model complexity isn't seen in the other models, suggesting further investigation is needed. Second, a KPI gap reduction between the Multi-scale Five model and baseline models was observed in the UNINA scenarios but not in the CARRS-Q scenarios. The expected performance drop from distractions affected the Multi-scale Five model the most. A proposed explanation is that this model better estimates the boundary between predictable and chaotic driver behaviour, which shifted in distraction scenarios. The theory remains untested, and since this model is excluded from final evaluations, future work is advised.

In the calibrated model results, four anomalies were identified. First, the Multi-scale Five model was excluded from the calibrated model results. This was a deliberate decision. Due to the nature and assumptions of the Multi-scale Five model it overfits for each individual trajectory and is incompatible with generalisation. Second, the TDIDM underperformed across all KPIs, contrary to the results in the original TDIDM paper Saifuzzaman et al. (2015). Differences in KPI metrics (SMAPE vs. RMSPE), calibration methodology (single vs. two-step), and software (PyGAD vs. Matlab GA) were examined. Repeating the test using two-step calibration and RMSPE showed that the differences persisted, mainly due to unusually high IDM errors in the original study. Third, the desired speeds of the Multi-scale model in the CARRS-Q scenarios are unrealistically high, around 90 to 120 km/h. This discrepancy may have resulted from task saturation effects reducing drivers personal preferences. When the, up to 60%, reduction is applied the speeds line up with the desired speed from IDM. This could be caused by two alternatives in the optimization process. Either the GA got stuck in a local minima or the high desired speed with reduction was the true global minima. Either way the impact on conclusions is limited. Finally, the spontaneous distraction in CARRS-Q scenario 3 showed unexpected task demand patterns, low during phone distraction onset and high during the conversation. The observed typical reaction from a driver to the start of the phone distraction is a short deceleration followed by a short acceleration. The model's limitation in reproducing the deceleration-acceleration reaction within a single interval leads to suboptimal fitting. This limits the Distraction model's accuracy and suggests the need for future studies on spontaneous distraction dynamics to refine the model's responsiveness.

5. Conclusion & Future Developments

The goal of this research was to develop, integrate and test a systematic distraction framework for traffic models, a domain untouched in existing literature. To this end a methodology was successfully defined and executed and the following answer was found.

The manner in which a framework which systematically defines distractions can be implemented and validated in a car-following model is by following the steps outlined in the methodology. This means first defining the logic and properties of the distraction framework. Subsequently an adequate TCI car-following model needs to be chosen and modified by implementing the framework. Lastly this Distraction model needs to be tested to prove that it is valid for the given distraction scope.

The paper in itself follows this guide and the results of the validation test for the built Distraction model show that these steps work and that the distraction framework provides adequate support to systematically model different types of distractions. To be more specific the results of the validation showed that the built Distraction model provided equivalent results to other models, which can support distractions and human factors, for at least two completely different types of distractions. This also means that a systematic approach of modelling distraction won't necessarily lead to an increase in performance for the chosen model. Instead it makes the model more flexible, easier to understand and simpler to expand upon.

As mentioned at the end of the above paragraphs the gap in existing literature has been filled by one solution. That does not preclude other solutions to fill this gap. Other manners of filling this gap are possible and might provide interesting pros and cons compared to the found method. To this end it is recommended that others research this topic of systematically modelling distractions in traffic models. This can be done either by improving upon this paper or by taking a completely new approach. In the case of wanting to continue improving upon this paper, the discussion contains multiple suggestions and weak points of this paper. By strengthening these weaknesses it should be possible to obtain better results. One recommendation which is not mentioned in the discussion is that the amount of validated building blocks/equations for the distraction trigger, intensity and effects are quite limited. This is due to the limited amount of distraction datasets available. By increasing the amount of building blocks it will become easier for others to use this framework for their own separate studies.

This work can also be used to further studies into driver behaviour when encountering distractions. This systematic approach to defining and modelling distractions should help to more systematically understand and analyse the behaviour of drivers.

Data Availability Statement All code used in this research is available at https://github.com/DwarfyAssassin/Distraction-framework. Do note that this repository doesn't include the data nor the results.

AI Tools Statement The following AI tools were used in the process of writing this paper. *LanguageTool* to check for spelling and grammar errors. *ChatGPT GPT-4.0* to help condense and shorten some sections of the thesis to fit the paper format.

References

Andersen, George J., and Craig W. Sauer. 2007. Optical information for car following: the driving by visual angle (dva) model. *Human Factors The Journal of the Human Factors and Ergonomics Society* 49, no. 5 (August 21, 2007): 878–896. https://doi.org/10.1518/001872007x230235. https://doi.org/10.1518/001872007x230235.

Calvert, Simeon C., Wouter J. Schakel, and J.W.C. Van Lint. 2020. A generic multi-scale framework for microscopic traffic simulation part ii – anticipation reliance as compensation mechanism for potential task overload. *Transportation Research Part B Methodological* 140 (August 9, 2020): 42–63. https://doi.org/10.1016/j.trb.2020.07.011. https://doi.org/10.1016/j.trb.2020.07.011.

Cuentas-Hernandez, Sandra, Xiaomeng Li, Mark J. King, and Oscar Oviedo-Trespalacios. 2023. The impact of road traffic context on secondary task engagement while driving. *Frontiers in Psychology* 14 (April 3, 2023). https://doi.org/10.3389/fpsyg.2023.1139373. https://doi.org/10.3389/fpsyg.2023.1139373.

Eltoweissy, Mohamed, Stephan Olariu, and Mohamed Younis. 2010. Towards autonomous vehicular clouds: a position paper (invited paper), 49:1–16. August 18, 2010. https://doi.org/10.1007/978-3-642-17994-5_1. https://doi.org/10.1007/978-3-642-17994-5_1.

Engström, Johan, Christopher A Monk, Richard J Hanowski, William J Horrey, John D Lee, Daniel V McGehee, Michael Regan, et al. 2013. *A conceptual framework and taxonomy for understanding and categorizing driver inattention.* 9984186943902771. Accessed February 15, 2025. https://iro.uiowa.edu/esploro/outputs/report/A-conceptual-framework-and-taxonomy-for/9984186943902771.

Fuller, R., C. McHugh, and S. Pender. 2007. Task difficulty and risk in the determination of driver behaviour. *European Review of Applied Psychology* 58, no. 1 (January 17, 2007): 13–21. https://doi.org/10.1016/j.erap.2005.07.004. https://doi.org/10.1016/j.erap.2005.07.004.

Fuller, Ray. 2011. *Driver control theory,* edited by Porter, 13–26. Elsevier Inc., January 1, 2011. https://doi.org/10.1016/b978-0-12-381984-0.10002-5. https://doi.org/10.1016/b978-0-12-381984-0.10002-5.

Gordon. 2008. *Crash studies of driver distraction,* 281–304. CRC Press, January 1, 2008. https://doi.org/10.1201/9781420007497.ch16. https://www.scopus.com/record/display.uri?eid=2-s2.0-85123140187&origin=inward&txGid=76a56180a28422b3cc8d8c76e172ea40.

Hamdar, Samer H., Hani S. Mahmassani, and Martin Treiber. 2015. From behavioral psychology to acceleration modeling: calibration, validation, and exploration of drivers' cognitive and safety parameters in a risk-taking environment. *Transportation Research Part B Methodological* 78 (May 18, 2015): 32–53. https://doi.org/10.1016/j.trb.2015.03.011. https://doi.org/10.1016/j.trb.2015.03.011.

Hamdar, Samer H., Martin Treiber, Hani S. Mahmassani, and Arne Kesting. 2008. Modeling driver behavior as sequential risk-taking task. *Transportation Research Record Journal of the Transportation Research Board* 2088, no. 1 (January 1, 2008): 208–217. https://doi.org/10.3141/2088-22. https://doi.org/10.3141/2088-22.

Irvine, Zhang, Miranda, García, Rodríguez, Saco, Koyal, et al. 2023. *Methodology and results: relevant use cases and safety-critical scenarios.* D1.6. July 27, 2023. https://i4driving.eu/deliverables/.

Kesting, Arne, and Martin Treiber. 2008. Calibrating car-following models by using trajectory data. *Transportation Research Record Journal of the Transportation Research Board* 2088, no. 1 (January 1, 2008): 148–156. https://doi.org/10.3141/2088-16. https://doi.org/10.3141/2088-16.

Li, Linbo, Yang Li, and Daiheng Ni. 2020. Incorporating human factors into lcm using fuzzy tci model. *Transportmetrica B Transport Dynamics* 9, no. 1 (October 24, 2020): 198–218. https://doi.org/10.1080/21680566.2020.1837033. https://doi.org/10.1080/21680566.2020.1837033.

Liang, Yulan, and John D. Lee. 2010. Combining cognitive and visual distraction: less than the sum of its parts. *Accident Analysis & Prevention* 42, no. 3 (April 12, 2010): 881–890. https://doi.org/10.1016/j.aap.2009.05.001. https://doi.org/10.1016/j.aap.2009.05.001.

Papathanasopoulou, Vasileia, and Constantinos Antoniou. 2015. Towards data-driven car-following models. *Transportation Research Part C Emerging Technologies* 55 (March 7, 2015): 496–509. https://doi.org/10.1016/j.trc.2015.02.016. https://doi.org/10.1016/j.trc.2015.02.016.

Punzo, Vincenzo, Biagio Ciuffo, and Marcello Montanino. 2012. Can results of car-following model calibration based on trajectory data be trusted? *Transportation Research Record Journal of the Transportation Research Board* 2315, no. 1 (January 1, 2012): 11–24. https://doi.org/10.3141/2315-02. https://doi.org/10.3141/2315-02.

Ranjitkar, Prakash, Takashi Nakatsuji, and Motoki Asano. 2004. Performance evaluation of microscopic traffic flow models with test track data. *Transportation Research Record Journal of the Transportation Research Board* 1876, no. 1 (January 1, 2004): 90–100. https://doi.org/10.3141/1876-10. https://doi.org/10.3141/1876-10.

Regan, Michael A., and Charlene Hallett. 2011. *Driver distraction: definition, mechanisms, effects, and mitigation,* 275–286. Elsevier Inc., January 1, 2011. https://doi.org/10.1016/b978-0-12-381984-0.10020-7. https://doi.org/10.1016/b978-0-12-381984-0.10020-7.

Saifuzzaman, Mohammad, and Zuduo Zheng. 2014. Incorporating human-factors in car-following models: a review of recent developments and research needs. *Transportation Research Part C Emerging Technologies* 48 (October 13, 2014): 379–403. https://doi.org/10.1016/j.trc.2014.09.008. https://doi.org/10.1016/j.trc.2014.09.008.

Saifuzzaman, Mohammad, Zuduo Zheng, Md. Mazharul Haque, and Simon Washington. 2015. Revisiting the task–capability interface model for incorporating human factors into car-following models. *Transportation Research Part B Methodological* 82 (October 22, 2015): 1–19. https://doi.org/10.1016/j.trb.2015.09.011. https://doi.org/10.1016/j.trb.2015.09.011.

———. 2017. Understanding the mechanism of traffic hysteresis and traffic oscillations through the change in task difficulty level. *Transportation Research Part B Methodological* 105 (October 28, 2017): 523–538. https://doi.org/10.1016/j.trb.2017.09.023. https://doi.org/10.1016/j.trb.2017.09.023.

Spall, James C. 2005. *Introduction to stochastic search and optimization: Estimation, simulation, and control.* John Wiley & Sons, March 11, 2005.

Tang, Quinteiro, Dopico, García, and Ahlström. 2023. *I4driving causal relationships between human/external factors and human drivers behaviors: modelling requirements and framework of testable hypotheses.* D1.5. October 26, 2023. https://i4driving.eu/deliverables/.

Van Lint, J.W.C., and S.C. Calvert. 2018. A generic multi-level framework for microscopic traffic simulation—theory and an example case in modelling driver distraction. *Transportation Research Part B Methodological* 117 (September 8, 2018): 63–86. https://doi.org/10.1016/j.trb.2018.08.009. https://doi.org/10.1016/j.trb.2018.08.009.

Victor, Engström, and Harbluk. 2008. *Distraction assessment methods based on visual behavior and event detection,* 135–161. CRC Press, October 15, 2008. https://books.google.nl/books?id=o7--7AS38tYC.

Wiedemann, Rainer. 1974. Simulation des straßenverkehrsflusses. Instituts Fur Verkehrswesen Der Universitat Karlsruhe, January 1, 1974. https://trid.trb.org/view/596235.

Young, Kristie L., Rachel Osborne, Sjaan Koppel, Judith L. Charlton, Raphael Grzebieta, Ann Williamson, Narelle Haworth, Jeremy Woolley, and Teresa Senserrick. 2019. What contextual and demographic factors predict drivers' decision to engage in secondary tasks? *IET Intelligent Transport Systems* 13, no. 8 (March 19, 2019): 1218–1223. https://doi.org/10.1049/iet-its.2018.5546. https://doi.org/10.1049/iet-its.2018.5546.

Young, Kristie Lee, Michael Arthur Regan, and John D Lee. 2008. *Factors moderating the impact of distraction on driving performance and safety,* 353–370. CRC Press, October 15, 2008. https://doi.org/10.1201/9781420007497-30. https://doi.org/10.1201/9781420007497-30.