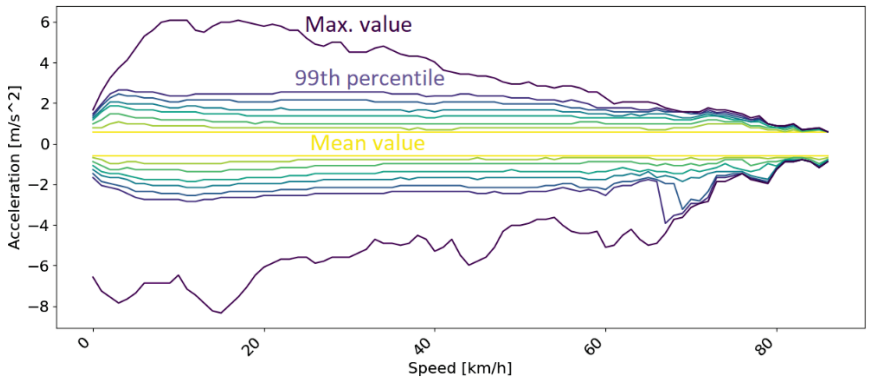
SURPRISE:

Using large scale fleet data to quantify the risk of being critically surprised in traffic

Public Report



Author: Mikael Ljung Aust, Malin Svärd, Jonas Bärgman
Date: 2025-09-29
Project within FFI Trafiksäkerhet och automatiserade fordon
Dnr: 2024-00823

FFI Fordonsstrategisk Forskning och Innovation
VINNOVA Energimyndigheten TRAFIKVERKET
FKG VOLVO SCANIA VOLVO

Contents

Summary and goals for the project

The next generation of cars will have unprecedented sensing capabilities both outside and inside the car and thus holds great potential in terms of being able to prevent traffic conflicts from becoming crashes. However, when looking at crash statistics, it is also very clear that the individual risk of experiencing a harmful crash is very low, compared to the total milage driven.

In other words, drivers are thus already very good at preventing traffic conflicts from becoming crashes. Also, market research shows that drivers perceive systems which give unnecessary warnings or interventions (in the sense that it's not perceived as safety critical) as untrustworthy, worthy of neither respect nor money. In fact, if the car brakes hard for no apparent reason, customers might demand to return the car and get their money back, because such behaviors can be almost scary. For future crash prevention systems to be successful, they therefore must be good, not only at predicting when drivers need help; but equally good at predicting when drivers do **not** need help. Otherwise, they risk being perceived as "cry wolf" systems and thus have their safety potential eradicated by poor reviews and low take rates.

So, how to formulate a design goal for future collision prevention systems that both prevents crashes and results in high acceptance from drivers? A good place to start is the basic realization that drivers generally do not intend to crash. Previous in-depth causation research at both Volvo and Chalmers (Ljung Aust, 2012) clearly shows that a crash typically comes as a *surprise* to the drivers(s). Simply put, sometimes when a traffic situation unfolds in a way that deviates from what a driver expects, there just isn't enough time left to resolve it once/if the driver becomes aware. One way to theoretically formulate the overall mission for future collision prevention systems is therefore to say that to prevent crashes from happening, we need to prevent driver surprise in critical situations.

Given that this might seem rather obvious, why is driver surprise not a dominating perspective in crash causation research? A key reason is a lack of the data needed to quantify the risk of driver surprise empirically and at scale. This is where the newly collected data set from Volvo Cars that will be used for this study provides a unique opportunity. However, to understand why that is; we first need to define what is empirically required to quantify the risk of driver surprise empirically.

First, we need kinematic information. If the driver is not on a conflicting trajectory with either another road user or something in the traffic environment, there is no conflict to prevent. Second, given that we should only warn or intervene on those rare occasions when the driver is unlikely to resolve the situation and would appreciate input from the vehicle, we need to be able to determine if the driver is aware of the unfolding conflict or not. If we cannot do that, preventing driver surprise becomes very difficult.

Third, the situation must also have a perceivable level of urgency to it, otherwise the driver will see no reason for the vehicle to act on it. Being able to limit warnings and interventions to situations of driver perceivable urgency is indeed what distinguishes a cry wolf system from an appreciated and useful one. For example, consider a lead vehicle cutting in at lower speed in front of an unaware driver. If this vehicle is 300 meters ahead, then even though the driver eventually would crash unless s/he brakes, the lead vehicle is so far away that we typically consider this just normal driving. Providing a warning at this point in time would therefore be the very definition of crying wolf.

To understand where to draw the line between urgent and non-urgent situations, we need to complement surprise with one more concept. Here, we believe *comfort* (or *comfort zone boundaries*, see Ljung Aust, 2012) to be the best candidate. To explain what it means, consider the following. When we learn how to drive, a large part of what we practice is how to maintain comfortable safety margins relative to other road users. When those margins shrink too much, we literally feel uncomfortable and do our best to restore them. For example, if a car starts following you very close in free-flowing traffic (e.g., at 3-4 meters behind at 70 kph), you will likely feel uncomfortable and wish for the following car to back off, even if you never have been in a situation where such behavior resulted in an actual crash.

Combining the concepts of *surprise* and *(dis)comfort* leads to an intuitive design goal for future collision prevention systems; they should help resolve traffic situations that might be both surprising and uncomfortable to the driver. Being alerted to something that is unpredicted and potentially dangerous is both objectively helpful (i.e., helps the driver resolve the situation safely) and likely to achieve high driver acceptance and usage.

Providing quantitative estimates of how surprising a situation will be and whether the driver would feel uncomfortable with it would thus provide an immensely useful tool for safety system developers, and the concrete problem becomes how to practically operationalize these estimates. Here, a fourth criteria presents itself. To unleash the full power of modern tools for developing computer models of the driver surprise and comfort zone boundaries, having only a few events in your data is not enough. Hundreds or thousands of events are needed.

Toward this fourth criteria and in terms of investigating driver surprise, the new dataset that Volvo has collected offers a new and possibly unique opportunities. It contains the necessary kinematic information described above, as well as the information on driver state required to analyze whether drivers are aware of a conflict situation or not. One goal of this project was therefore to leverage this large naturalistic driving data set to investigate data driven methods for classification and quantification of surprise.

For the fourth criteria and comfort zone boundaries, a more directed data set (i.e. where drivers can be explicitly asked about their comfort level as different scenarios play out) than fleet data is needed. To achieve the necessary data set size, the project therefore decided to investigate if crowd sourced data can be used to collect data to quantify driver comfort zone boundaries.

It should be noted that being able to predict surprise and understand comfort zones is important also for automated driving. Understanding how and when one might generate surprise or violate the comfort zones of other road users is key to providing predictable behavior in automated vehicles, which often is pointed to as a key to public acceptance and usage. This project thus supports the introduction and deployment of automated driving as well.

Background

The concept of surprise in an automative setting has recently gotten quite some attention through Waymo publications on the topic (Dinparastdjadid, Supeene, & Engstrom, 2023; Engström, Liu, Dinparastdjadid, & Simoiu, 2023); they are pursuing surprise as a way to establish an objective way to quantify driver response timing, supporting the development and assessment of their “computer-driver”.

A few years back, the predictive processing framework proposed by Engström et al. (2018) emphasized the role of rare events and surprise in crash causation and modelling of human behavior. Other work related to the concept of surprise include expectation mismatch research (Bianchi Piccinini, Engström, Bärgman, & Wang, 2017; Eiríksdóttir, Engström, & Bärgman, 2017; Räsänen & Summala, 1998), research into driver response processes (Morando, Victor, Bengler, & Dozza, 2020) and “oops” reactions (Dozza & González, 2013; McGehee & Carsten, 2010). Also recent work on higher levels of automation has acknowledge the importance of expectations and surprise (Victor et al., 2018).

As argued above, surprise needs to be complemented by the concept of driver (dis)comfort to provide a full picture. Here the earliest reference of comfort in driving dates back to Gibson and Crooks (1938), that proposed a “field of safe travel” to explain how driver manage the driving task. A more recent as well as influential work on comfort was presented in Summala (2007).

Approximately ten years ago, VCC and Chalmers were key partners in an FFI project (DCBIN; Vinnova, 2016) that explicitly sought to quantify comfort zone boundaries (CZBs) in a turning scenario (Bärgman, Smith, & Werneke, 2015). Since then, research on CZB in the automotive domain include the proposal of a forward collision warning algorithm based on CZB (Kovaceva, Bärgman, & Dozza, 2022) and the use of CZB to establish reference driver models for vehicles with higher levels of automation (Olleja, Rasch, & Bärgman, 2023).

However, despite substantial research efforts around the concepts of surprise and comfort, the idea of considering these two concepts in combination is a novel approach (as far as the authors are aware).

Aim, research questions and method

The project has consisted of two main work areas. One focused on investigating which methods might be most suitable for researching parameters that can be used to quantify surprise in large naturalistic driving sets. The second focused on which methods might be most suitable if one wants to estimate driver comfort zones and generate large data sets.

For the first area, the project has made use of a large naturalistic dataset collected by Volvo Cars as part of its sensor and active safety function development. This dataset contains thousands of hours of driving and is structured to be largely representative of weather, light, and traffic environment conditions around the world. What makes this dataset uniquely suitable for the project is that it contains both continuous gaze direction data and precise mapping of nearby traffic participants.

For the second area, the possibilities to source larger amounts of data than normally obtainable in local studies using the mechanism of crowd-sourcing has been explored. Below, the work done and results achieved in each area will be described in more detail.

Parameters that quantify surprise in fleet data

To narrow the scope for the pilot study to something manageable, it was decided that the main aim for this part of the project would be to study three very specific aspect of how driving context may influence driver behaviors and the underlying risk of being surprised. The data set used as basis for the analyses comes from a large data collection effort by Volvo during 2021-2025. Data was collected by Volvo car's test drivers on field excursions over large parts of Europe, the US, and some parts of Eastern Asia (Japan, Malaysia). It consists in total of approximately 15000 hours of manual driving, of which a subset was used in this project.

Glance duration as a function of time headway

It can be hypothesised that drivers generally adapt their off-road glance behavior as a function of contextual cues as time headway (THW) or time-to-collision (TTC). To confirm this, we first studied the duration of off-road glances in relatively stationary car-following scenarios on highways. Particularly, off-road glance data were extracted from driving segments with stable THWs. The THWs were binned into discrete intervals and the off-road glance durations were matched to each bin accordingly.

It was observed that the occurrence of long glances was less in segments with short THW than in segments with long THW. Consequently, we concluded that drivers adapt their glances to the higher risk (smaller error margins) in these scenarios, thus decreasing the likelihood of not detecting a surprising maneuver by the car in front. The findings from this study will be described in a scientific journal publication that currently is in preparation.

Glance behavior as a function of inverse time-to-collision

Another important aspect of driving context during car-following is the (inverse) TTC to the car in front. Similarly to the THW study, glance duration was analysed and sorted into corresponding inverse TTC intervals. However, the main point of interest was to see how the tendency to initiate a glance changed with the inverse TTC.

Previous research (Markkula, G., et al., 2016) has found that drivers who are in a critical situation and look up at an inverse TTC >= 0.2 s^{-1} relative to the car in front immediately starts to brake. A natural consequence of this would be that drivers generally do not initiate off-road glances at that inverse TTC (or higher inverse TTCs). However, this has not previously been confirmed in everyday driving data, such as the data used in this project. As expected, the analysis showed a decrease in both glance frequency (number of glances per time sample) and duration with increasing inverse TTC values, with almost no glances initiated at an inverse TTC value above 0.3 s^{-1}. It was concluded that drivers in everyday (non-critical) car-following situations may look away from the road even at inverse TTC >= 0.2, but that most drivers refrain from initiating off-road glances at an inverse TTC between 0.2 and 0.3 s^{-1}.

As for the glance durations, the mean values were approximately constant across the entire range of inverse TTC values. However, the decrease in glance duration as a function of increased inverse TTC became increasingly apparent at higher percentiles, with the 95th and 99th duration percentiles corresponding to much longer duration (~1-1.7 s, equivalent to an increase in 200-300%) when initiating the glances at low inverse TTCs compared to at high inverse TTCs (see Figure 1 for a density plot of glance duration as a function of inverse TTC). The findings from this study are also in preparation for being submitted for publication in a scientific journal.

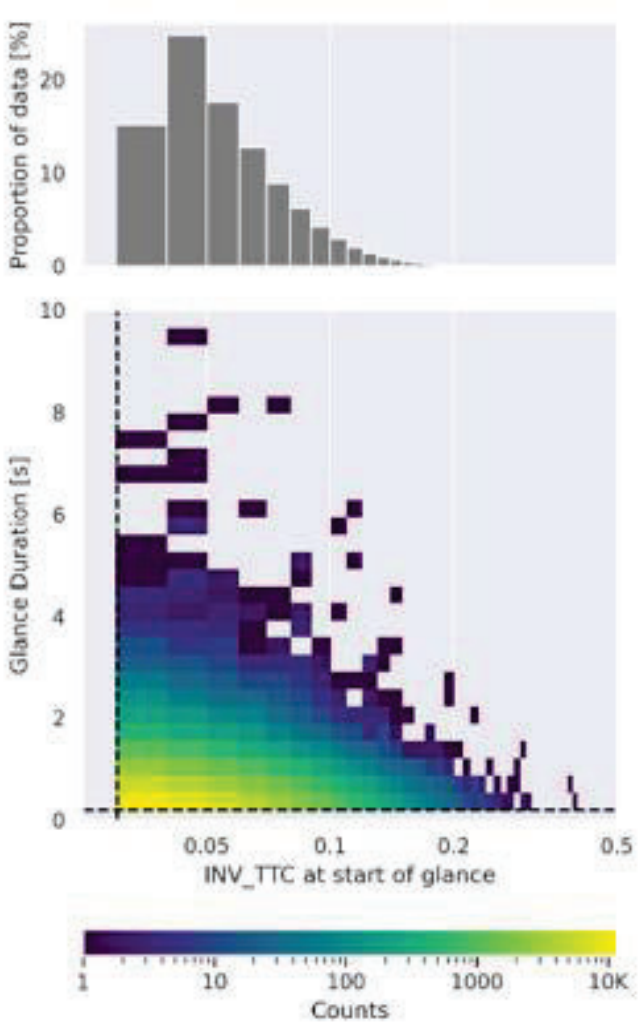


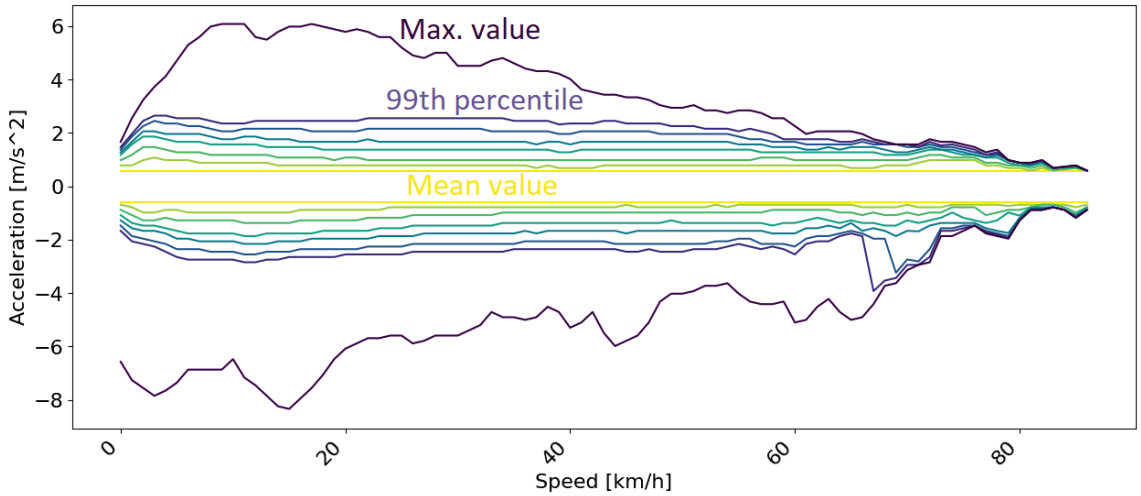
Figure 1 Density plot of glance duration and inverse TTC at glance initiation

Longitudinal speed as a function of longitudinal acceleration

The final analysis on the Volvo cars data set aimed to investigate how the longitudinal acceleration range used in everyday-driving changes depending on current longitudinal speed. Here, we wanted the analysis to cover wide range of traffic environments, and, thus, no pre-filtering was performed on the data. This kind of analysis serves to inform about the overall comfort zone boundaries of the current driver population, which can be used to adapt AD/ADAS functionality.

During the analysis, the longitudinal speed was binned into 1km/h wide bins, each in which a gamma distribution was fit to the corresponding observed acceleration values. Acceleration values within -0.5 to 0.5 m/s^2 were discarded to avoid studying minor adjustments at "constant" speed. Surprisingly, it was observed that the preferred acceleration values were very similar over the entire speed range (0-100 km/h). 99th percents of all collected time samples corresponded to an acceleration in the range -3 to 2 m/s^2. However, the maximum observed values were substantially higher, and the speed-acceleration envelope seemed to follow a gamma distribution for both positive and negative accelerations. Thus, the difference between the maximum acceleration and the 99th percentile acceleration value in each speed

Figure 2: 50th to 100th percentiles of positive and negative acceleration, as a function of speed. Note that accelerations with absolute values < 0.5 m/s2 were excluded from the data set.



bin decreased with increasing speed. As shown in Figure 2, the observed differences between higher percentiles were almost negligible above 70 km/h, while above 80 km/h, all acceleration percentile values remained very close to the mean accelerations.

While the results above are both limited and preliminary, they were promising enough to warrant continued development in the follow-up project, QUADRARUM.

Crowd-sourcing experiment pilot for establishing perceived safety and comfort zone boundaries

The project application summary states:

> “This project uses the concepts of "surprise" and "comfort" to formulate the vision for future collision avoidance systems. Essentially, the goal is to intervene only in traffic situations that are both surprising and uncomfortable from the driver's perspective. If successful, this approach leads to both fewer accidents and greater acceptance of warnings and interventions.”

In this project, we have explored the issue from multiple angles, with a primary focus on developing and testing a method to experimentally identify the boundary—both in time and space—between comfortable and uncomfortable traffic situations. We refer to these boundaries as Comfort Zone Boundaries (CZB), a concept closely related to perceived safety, but operationalized as a discrete boundary rather than a continuous measure.
To assess the feasibility of this method, we conducted several small-scale pilot studies and one larger study. While the results are preliminary, they are promising enough to warrant continued development in the follow-up project, QUADRARUM.

The experimental design process began with a literature review to update our understanding of existing methods for identifying CZBs. As no major advancements had occurred in the field since our initial project proposal, we opted to test a modified version of the method developed by Xiaolin He at TU Delft (He, 2024).

We selected four distinct traffic scenarios (see Figure 3), and for each scenario, we generated a series of video clips using the CARLA simulation software. These clips varied in terms of criticality and spatiotemporal timing. Participants viewed a selection of these clips and responded to a set of questions designed to assess their perception of safety and comfort.

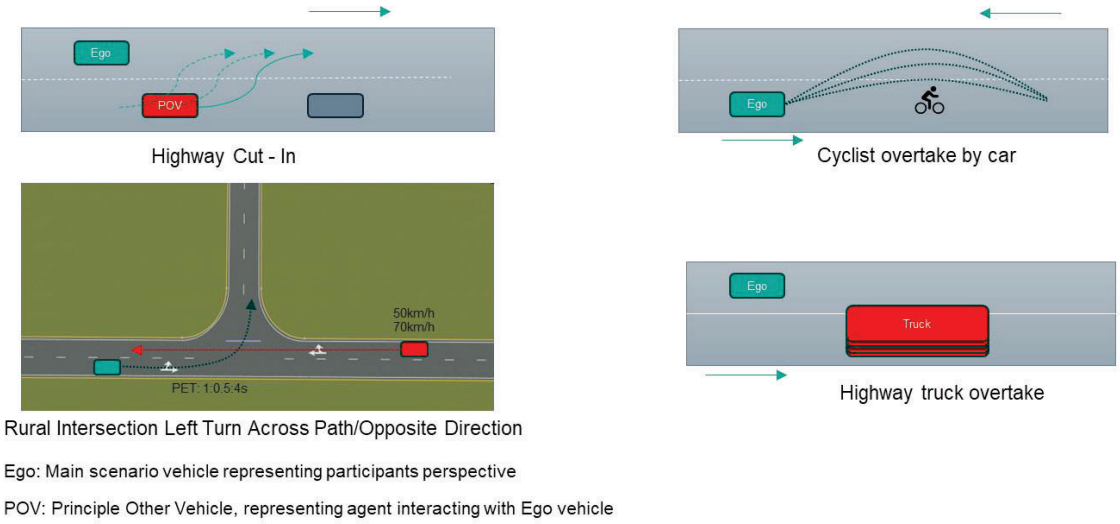


Figure 3: An illustration of the four scenarios in the larger pilot study.

The experiment was conducted via a crowdsourcing platform. Participants registered through the service and were redirected to the study interface, where they completed two sessions of approximately one hour each, spaced at least one day apart. During the sessions, participants provided demographic and background information, viewed the scenarios in randomized order, and answered questions after each clip (see Figure 4).

Compared to the original study by He et al. (2025), we refined the design to more precisely capture the spatial and temporal dimensions of perceived safety and CZBs. Specifically, we reduced the interval between video clips to a few hundred milliseconds (but with overlaps), whereas He used 6-second clips without overlap. A total of 33 participants completed the full

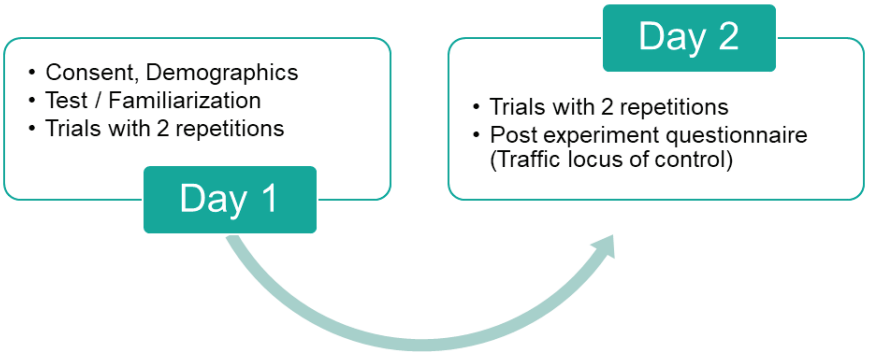


Figure 4: An illustration of the study setup

final study and provided usable data. Since the use of crowdsourcing for data collection was new to the project team, we adopted an iterative approach to explore its potential within the project. We conducted several smaller studies to familiarize ourselves with the procedures and optimize the setup. During this process, we refined our understanding of the experiment tool and the dynamics of working with crowdsourced participants, which led to the exclusion of some data along the way. We also collaborated closely with the provider of the experiment platform to ensure the system functioned as intended and met our specific requirements.

The data analysis is still in its early stages and will continue in the follow-up project, QUADRARUM, where we plan to conduct additional studies using the same method—refined based on insights gained from the pilot studies.

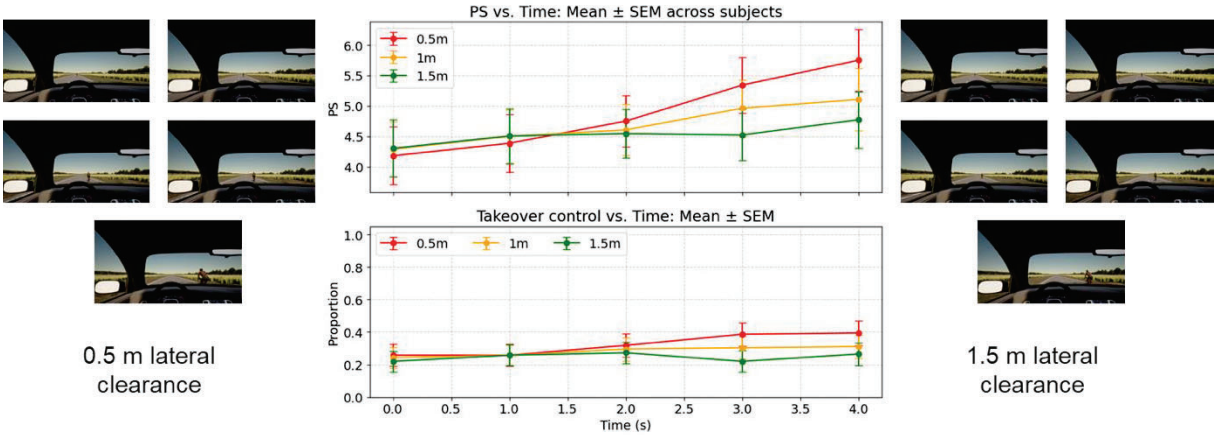


Figure 5: En example illustration of results from the crowd source study of a car overtaking a cyclist.

Figures 5 and 6 present two examples of preliminary results. Figure 5 includes illustrations of the two extremes in terms of levels of criticality in the cyclist overtaking scenario. The results clearly demonstrate differences in perceived risk depending on the lateral clearance

between the vehicle and the cyclist during overtaking. Notably, perceived risk increases as the clearance decreases, with the most pronounced differences observed when the vehicle passes very close to the cyclist.

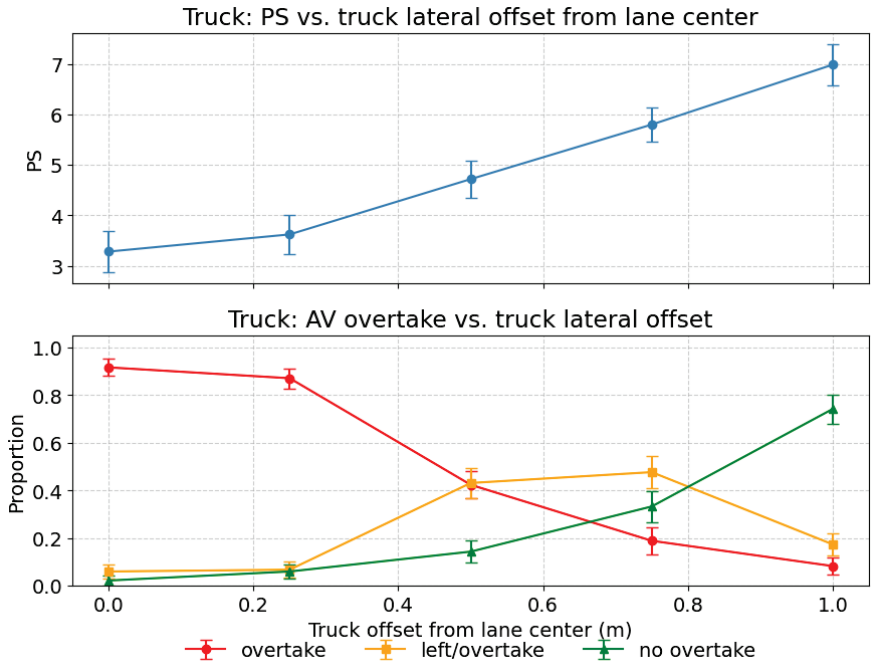


Figure 6: Results of the study of drivers approaching a truck with different lateral position in relation to the lane center.

Figure 6 presents results from the study on perceived safety and comfort zone boundaries in the context of a car overtaking a truck. The lower panel of the figure clearly illustrates how participants' expectations of an automated vehicle's behavior—specifically whether it should overtake, shift further to the left before overtaking, or refrain from overtaking altogether—varied depending on both the criticality of the situation and the proximity to the truck. These findings suggest that perceived safety and behavioral expectations are closely linked to spatial and temporal factors in overtaking scenarios, and that it is possible to quantify the difference spatiotemporally.

In summary, the development and initial assessment of a method to quantitatively establish perceived safety and comfort zone boundaries (CZB) has been successful. While we have identified several aspects of the study design that can be improved, the progress made is in line with expectations for a pilot study. Building on these insights, we will continue refining and applying the method in the follow-up project, QUADRARUM.

Goal fulfillment

Looking at the overall results, we think that this project has fulfilled its goals. Starting with fleet data; while it provides completely new opportunities to follow the course of events that precede an accident or near-eye incident with very high accuracy, it also places high demands on the ability to both handle and analyse very large amounts of data. In this project, we have successfully tried several analyses that follow from the issues described in the project application, in terms of how driving situations develop. It also turned out to be feasible to create a meaningful design for crowd sourced studies of comfort zone boundaries in a way that allows for collection of relatively large datasets at a speed not really available for traditional local user studies.

Given that this work has been successful, we have therefore applied for, and been granted funding for, a significantly larger follow-up project (QUADRARUM, diary no. 2025-00834) that intends to tackle the same issues more deeply and over a longer period. This project has just started; The doctoral student recruitment is completed, and we have had a first promising kick-off with our Scientific Advisory Board that has helped us sharpen the research questions even further.

A somewhat unexpected side effect of the project is that one of the fleet data analyses turned out to be just what was needed to close a previously challenging gap in the safety argumentation that Volvo Cars is building up to show that the self-driving vehicles we are developing will behave safely. So the pilot study actually has contributed directly to the development of self-driving cars.

For Volvo, this is in many ways the first step into a new kind of research and development environment where in many cases engineering judgement can be improved or replaced with relevant empirical data from the field. For safety systems, this means a completely new approach and way of working for functional development. Instead of having to make an assessment to the best of your ability, you can study both the conditions for success and the effects of introduced measures directly in the car fleet. The importance of this step forward is hard to overestimate.

Publications

This pilot study has resulted in two publications that currently are in preparation but not yet submitted.

Participating organizations and contact persons

Chalmers University
Contact: Jonas Bärgman,
jonas.bargman@chalmers.se

Volvo Cars
Contact: Mikael Ljung Aust,
mikael.ljung.aust@volvocars.com

References

Bianchi Piccinini, G., Engström, J., Bärgman, J., & Wang, X. (2017). Factors contributing to commercial vehicle rear-end conflicts in China: A study using on-board event data recorders. *Journal of Safety Research, 62*, 143-153. doi:https://doi.org/10.1016/j.jsr.2017.06.004

Bärgman, J., Smith, K., & Werneke, J. (2015). Quantifying drivers' comfort-zone and dread-zone boundaries in left turn across path/opposite direction (LTAP/OD) scenarios. *Transportation Research Part F: Traffic Psychology and Behaviour, 35*, 170-184. doi:10.1016/j.trf.2015.10.003

Cordellieri, P., Baralla, F., Ferlazzo, F., Sgalla, R., Piccardi, L., & Giannini, A. (2016). Gender Effects in Young Road Users on Road Safety Attitudes, Behaviors and Risk Perception. Frontiers in Psychology, 7. https://doi.org/10.3389/fpsyg.2016.01412

Dinparastdjadid, A., Supeene, I., & Engstrom, J. (2023). Measuring Surprise in the Wild. *Arxiv*. doi:https://doi.org/10.48550/arXiv.2305.07733

Dozza, M., & González, N. P. (2013). Recognising safety critical events: Can automatic video processing improve naturalistic data analyses? *Accident Analysis & Prevention, 60*, 298-304. doi:https://doi.org/10.1016/j.aap.2013.02.014

Eiríksdóttir, H. H., Engström, J., & Bärgman, J. (2017). *The perfect mismatch: Analysis of rear-end crash causation mechanisms based on naturalistic crash data*. Paper presented at the 6th International Symposium on Naturalistic Driving Research, The Hague, The Netherlands.

Engström, J., Bärgman, J., Nilsson, D., Seppelt, B., Markkula, G., Piccinini, G. B., & Victor, T. (2018). Great expectations: a predictive processing account of automobile driving. *Theoretical Issues in Ergonomics Science, 19*(2), 156-194. doi:10.1080/1463922X.2017.1306148

Engström, J., Liu, S.-Y., Dinparastdjadid, A., & Simoiu, C. (2023). Modeling road user response timing in naturalistic settings: a surprise-based framework. *Arxiv*. doi: https://doi.org/10.48550/arXiv.2208.08651

Gibson, J. J., & Crooks, L. E. (1938). A theoretical field-analysis of automobile driving. *The American Journal of Psychology, 51*(3), 453-471. doi:10.2307/1416145

He, X. (2024). Driver's perceived risk in relation to automated vehicle behaviour TU Delft]. The Netherlands. http://resolver.tudelft.nl/uuid:14f15a1f-7f81-43b4-9d4c-8a28c4e683cf

Kovaceva, J., Bärgman, J., & Dozza, M. (2022). On the importance of driver models for the development and assessment of active safety: A new collision warning system to make overtaking cyclists safer. Accident Analysis & Prevention, 165, 106513. doi:https://doi.org/10.1016/j.aap.2021.106513

Ljung Aust, M. (2012). Improving the Evaluation Process for Active Safety Functions. Dissertation, Chalmers University of Technology. ISBN 978-91-7385- 663-8.

Markkula, G., Engström, J., Lodin, J., Bärgman, J., & Victor, T. (2016). A farewell to brake reaction times? Kinematics-dependent brake response in naturalistic rear-end emergencies. Accident Analysis & Prevention, 95, 209–226. https://doi.org/10.1016/j.aap.2016.07.007McGehee, D. V., & Carsten, O. (2010). Perception and biodynamics in unalerted precrash response Annals of advances in automotive medicine, 54, 315-332.

Morando, A., Victor, T., Bengler, K., & Dozza, M. (2020). Users' Response to Critical Situations in Automated Driving: Rear-Ends, Sideswipes, and False Warnings. IEEE Transactions on Intelligent Transportation Systems, 1-14. doi:10.1109/TITS.2020.2975429

Olleja, P., Rasch, A., & Bärgman, J. (2023, October 26–27). A reference driver model for overtaking a cyclist. Paper presented at the ICTCT, Catania, Italy

Rhodes, N., & Pivik, K. (2011). Age and gender differences in risky driving: the roles of positive affect and risk perception.. Accident; analysis and prevention, 43 3, 923-31 . https://doi.org/10.1016/j.aap.2010.11.015.

Räsänen, M., & Summala, H. (1998). Attention and expectation problems in bicycle–car collisions: an in-depth study. Accident Analysis & Prevention, 30(5), 657-666. doi:10.1016/S0001-4575(98)00007-4

Summala, H. (2007). Towards understanding motivational and emotional factors in driver behavior: Comfort through satisficing. In C. P. Cacciabue (Ed.), Modelling driver behavior in automotive environments (pp. 189-207). London, UK: Springer.

Victor, T., Tivesten, E., Gustavsson, P., Johansson, J., Sangberg, F., & Ljung Aust, M. (2018). Automation Expectation Mismatch: Incorrect Prediction Despite Eyes on Threat and Hands on Wheel. Human Factors, 60(8), 1095-1116. doi:10.1177/0018720818788164

Vinnova. (2016). Gränser för förartrygghet i korsningar (DCBIN) - an FFI project. Retrieved from https://www.vinnova.se/p/granser-for-forartrygghet-i-korsningar-dcbin/