Visual Sampling Behavior Does not Explain Risk Perception: A Data-Driven xAI Investigation

Martin Lorenz
lorenz@cs.uni-leipzig.de
ScaDS.AI, Leipzig University
Leipzig, Germany

Jan Hilbert
jan.hilbert@uni-leipzig.de
ScaDS.AI, Leipzig University
Leipzig, Germany

Philipp Michael Markus Peter Asteriou
philipp.asteriou@it-u.at
Interdisciplinary Transformation University
Linz, Austria

Philipp Wintersberger
philipp.wintersberger@it-u.at
Interdisciplinary Transformation University
Linz, Austria

Patrick Ebel
ebel@uni-leipzig.de
ScaDS.AI, Leipzig University
Leipzig, Germany

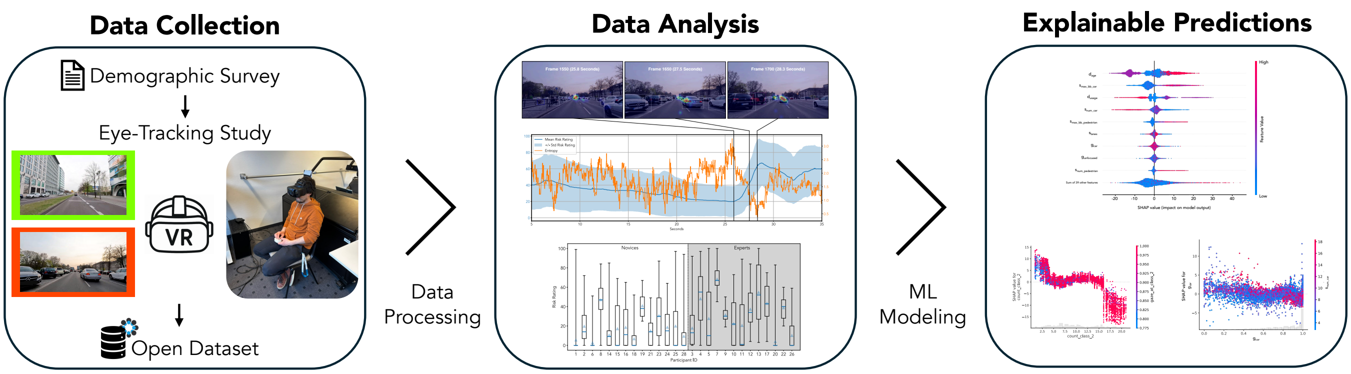


Figure 1: Participants watched dashcam videos in VR and continuously rated perceived risk (left). After preprocessing and analysis (middle), we trained a ML model to generate explainable predictions from video and eye-tracking data (right).

Abstract

How do drivers perceive risk? Understanding what situations and factors cause drivers to perceive situations as critical can improve our understanding of road user behavior and inform automated driving technology. To investigate the factors that shape drivers' risk perception, we conducted an eye-tracking study with 27 participants who watched dashcam videos and continuously rated the perceived risk of various driving situations. Using the resulting dataset, we developed a computer vision-based machine learning approach that generates explainable predictions of perceived risk from video and eye-tracking data. Our SHAP analysis reveals that the proximity of objects and number of cars in a scene are the most significant contributors to perceived risk. Most interestingly, while people tend to sample similar objects in critical situations, their risk perception remains highly personal making visual sampling behavior a weak predictor of perceived risk. Overall, our explanations reveal non-linear insights beyond previous work, suggesting that risk perception is not only shaped by visual input, but primarily by cognitive processes which is in line with theoretical models of situation awareness. The dataset, source code, and a comprehensive usage guide are publicly available.

CCS Concepts

• **Human-centered computing → HCI theory, concepts and models**; **Empirical studies in HCI**; **User models**.

Keywords

Risk Perception, Situation Awareness, Driving Simulator, Machine Learning, Explainable AI, Computational Modeling

ACM Reference Format:

Martin Lorenz, Jan Hilbert, Philipp Michael Markus Peter Asteriou, Philipp Wintersberger, and Patrick Ebel. 2025. Visual Sampling Behavior Does not Explain Risk Perception: A Data-Driven xAI Investigation. In *17th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '25), September 21–25, 2025, Brisbane, QLD, Australia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3744333.3747810

AutomotiveUI '25, Brisbane, QLD, Australia

ACM ISBN 979-8-4007-2013-0/2025/09
https://doi.org/10.1145/3744333.3747810

1 Introduction

What shapes drivers' risk perception so that they assess situations appropriately and react accordingly in case an intervention is needed? Investigating this question is key to understanding driver behavior in dynamic and often ambiguous traffic environments. Risk perception plays a critical role in road safety, driver decision-making, and in the broader context of automated driving. Understanding which situations drivers perceive as critical – and why – can inform traffic psychology, the design of next-generation driver assistance systems, and Automated Vehicles (AVs).

In a perfect world, drivers would constantly monitor the driving environment and calibrate their subjective risk in a way that it matches the objective risk of future events. However, questions of environment monitoring, situation awareness, and risk are complex. First, it has been shown that drivers differ in their understanding of driving situations and even in what defines a situation or context change [2]. Second, even attentive drivers who appropriately monitor the road environment do not always react appropriately to hazards [51]. Third, perceived risk consists of epistemic and non-epistemic components [22], leading to differences in risk perception between drivers based on their experience or ability [8]. For example, an experienced race driver may be able to successfully navigate a situation that a novice driver would consider a dead end. Consequently, the chain of thought between visual perception and potentially successful reactions needs to be researched more thoroughly. In this study, we focus on the initial stage of this process by examining what factors influence drivers' subjective risk assessment.

We argue that understanding and predicting perceived risk is important along three dimensions. First, it can help AVs determine whether drivers assess situations similarly as AV algorithms. In SAE level 2 driving [1], this information can be utilized to inform driver state assessment and monitoring support systems. In level 3, it can help to determine whether a driving situation is suitable for handing back control to the human driver. These take-over requests are a critical component in level 3 automated driving and represent a general problem in the interaction between humans and automation that has yet to be fully resolved [11]. Second, knowledge of perceived risk can guide AV behavior to align with user expectations, such as adjusting driving behavior in situations that are likely to be experienced as critical, thereby improving comfort and trust [43]. Third, the ability to predict perceived risk can support the generation of explanations in AVs, helping automated driving systems determine when and why to deliver information that improves user understanding and trust in automation [38, 55]. Considering these dimensions, it becomes clear that being able to predict and explain what drivers perceive as risky is not only beneficial for traffic psychology, but also necessary to develop vehicles that are aware and adaptive of how humans assess traffic situations.

Despite its significance, the question of how risk perception unfolds in real-world traffic scenarios remains underexplored, particularly in dynamic contexts [36]. In this paper, we address this gap by studying perceived risk as a continuous, temporally unfolding phenomenon, integrating eye-tracking data and computer vision approaches with explainable Machine Learning (ML). We conducted a user study in Virtual Reality (VR) with 27 participants, who watched dashcam videos of various urban traffic scenarios and continuously rated their perceived risk using a self-developed rotary controller. Throughout the study, we recorded participants' gaze behavior. To enhance the data we employ YOLO [41] for object detection enabling us to analyze which objects in the environment the participants focused on. Based on this multimodal data, we propose a ML model that predicts drivers' perceived risk based on features of the traffic scene, demographic information, and participants' visual attention allocation. To interpret the model's predictions, we apply SHapley Additive exPlanation (SHAP), an explainable AI method, grounded in game theory, that provides explanations of feature contributions. This allows us to generate insights into how factors such as object proximity or the number of surrounding vehicles or pedestrians influence perceived risk.

While prior research on risk perception has often leveraged eye-tracking data, these studies predominantly focus on low-level eye movement metrics (e.g., saccade durations or fixation counts) within binary hazard detection tasks [32, 33]. However, they overlook a crucial question: *what* do drivers actually look at, and how does their visual sampling behavior shape their perception of risk? To answer this question, we investigate risk perception through the lens of level 1 perception in Endsley's Situation Awareness model [20] – that is, the identification of relevant objects in the environment: *Is visually sampling specific objects predictive of perceived risk?* Furthermore, we go beyond prior work and binary classification approaches by adopting a continuous, real-time measure of perceived risk. This aligns with calls from prior work (e.g., Asteriou et al. [2], Moran et al. [36]) and enables us to derive more fine-grained insights into the dynamics and individual differences of risk perception, which is relevant for automated driving technology. Furthermore, we examine risk perception across a broader and more ecologically valid set of traffic scenarios, encompassing both ambiguous and unambiguous driving situations. Finally, while recent work such as De Winter et al. [10] uses static images to predict perceived risk, we incorporate the dynamic context of driving videos, combining eye-tracking with computer vision and explainable AI techniques to model and interpret perceived risk unfolding over time.

Our results show that information about what drivers are looking at is a weak predictor of perceived risk. Combining our modeling results with the descriptive results obtained from the analysis of the collected datasets, our findings indicate that while people are consistent in *perception* (level 1 SA according to [20]), they differ widely in *comprehension* (level 2 SA) and *projection* (level 3 SA). This makes Situation Awareness (SA) and risk perception in traffic an individual, experiential matter. This finding is underscored by the fact that contextual information about the driving scene and individual differences are the most relevant contributors to our ML predictions. The latter is consistent with previous work showing that risk perception is highly personal and, for example, differs significantly between novices and experts [8]. In summary this paper makes the following **contributions**:

- **An openly available and easily extensible eye-tracking dataset:** We present and share a dataset of perceived risk ratings of drivers (N=27) watching dashcam videos of urban traffic. The dataset consists of participants' *demographic* *data*, such as age and frequency of car use, the detected YOLO bounding boxes for each frame, and the recorded eye-tracking data for each participant. The dataset is accompanied by extensive documentation to facilitate its use and extension.

- **An explainable ML approach for perceived risk prediction:** We present an xAI approach that can predict drivers' perceived risk based on scene features, demographic information, and gaze behavior. The approach generates not only predictions, but also explanations of perceived risk that capture non-linear dependencies between features.
- **Insights into perceived risk perception connected to SA theory:** We present an analysis of our dataset and modeling results through the lens of Endsley's SA model [20], revealing that perceived risk is not solely determined by where people look, but rather by how they interpret the entire scene as a whole based on their experience.

2 Related Work

Understanding how drivers perceive risk requires engaging with different strands of research, such as SA, models of attention allocation, and recent work on risk perception and prediction. Below, we review the relevant findings of these areas and analyze how our work contributes to them using ML.

2.1 Situation Awareness and its Components

Situation Awareness (SA) is a theoretical framework often used in transportation studies on risk perception [36] or supervisory control [51]. SA is defined as the *"perception of the elements in the environment, within a volume of time and space, the comprehension of their meaning, and projection of their status into the near future"* [21]. According to Endsley [21], there are three levels of SA, namely *perception, comprehension*, and *projection*. A lot of research so far has emphasized the level of *perception* as it is less abstract than the other levels and, thus, more operationalizable. Zhang et al. [58] reviewed 25 articles on SA measurements, where eye-tracking was the most widely used one. Other methods, such as the situation awareness global assessment (SAGAT) [21] or the situation awareness rating technique (SART) [6] either strictly distinguish between the levels and consider the concept as a whole or are highly subjective.

In the context of driving, a great variety of factors influence SA. A study by Yang et al. [57] lists road (traffic and road conditions) and driver characteristics (gender, age, experience), driver states (emotions, fatigue), distractors, cognitive abilities, and properties of the environment on the three levels of perception (other entities, signs, hazards), understanding (locations and speeds of traffic objects), and prediction (behavior of the ego vehicle and other road users). By structural equation modeling, they showed that drivers' cognitive abilities had the highest influence on SA, followed by driver states and driver characteristics.

Another important theoretical perspective on SA is the Salience, Effort, Expectancy, and Value (SEEV) model [52]. This model describes how operators allocate visual attention in dynamic environments. It predicts where people will look at, given the visual salience of objects in a scene, the expectancy of changes in certain Area of Interests (AOIs), the effort it takes to shift the gaze from the current point to another AOI, and the value an AOI has for the task. For individual risk prediction, expectancy and value are of special interest, as these factors might not be objectively assessable. SEEV is increasingly used in driving research. Scharfe-Scherf et al. [44] present a cognitive model that directs attention to different areas of interest in take-over scenarios, and the authors suggest addressing eye-tracking in the future. Another work by Du et al. [14] combines SEEV with ACT-R and suggests investigating more real-life situations with diverse participants.

2.2 Perceived Risk: Psychology and Prediction

Risk perception has been extensively studied from a behavioral perspective. In one of the fundamental works, Slovic [47] distinguishes between risk assessment (analytic evaluation of risk) and risk perceptions (fast, intuitive judgments), where one's expertise and experience toward particular situations help to avoid biases and misconceptions. Li et al. [26] discuss subjective risk perception based on previous works by Stuck et al. [48] and Numan [37]: Risk perception involves the probability and severity of potential accidents, including both a relational (based on previous experiences and knowledge) and a situational (context-specific belief that a situation yields negative outcomes) component.

In the past, risk perception was frequently investigated when researching explanations for higher crash rates among young drivers, which have shown (at least partly) to stem from them systematically underestimating risk in traffic [23, 50]. In addition to age, driving experience also plays a significant role. According to Crundall [8], novices perform poorly when predicting hazards, while experts have more attentional resources and react to cues earlier [9]. This was also confirmed by Pradhan et al. [40] , who found that inexperienced drivers were less likely to fixate on potential hazards than experienced drivers. In addition, Moran et al. [36] found in a review on hazard perception that age and experience are the most influential features for perceived risk.

More recently, perceived-risk has increasingly been addressed in automated driving studies on trust in AVs [26], for example, to adjust driving styles. Another line of research addresses driver vigilance and driver monitoring systems. Arguably, AVs need to predict whether a driver monitors the road environment (on level 2) or if they would be able to take back vehicle control in case of a take-over request (level 3). De Winter et al. [10] summarized various works and listed factors that influence perceived risk, such as visibility, weather, headway, lane width, or proximity to other road users. In their work, they also proposed to focus more on predicting than merely measuring the concept, performed an image-based online study, and conducted a regression analysis to predict perceived risk. Using dashcam videos, Bazilinskyy et al. [4] assessed risk perception in different countries and found that respondents were better in assessing the risk of scenarios in their own cultures. Later, Driessen et al. [13] showed that GPT-4V was able to generate risk ratings that highly correlated with human study participants data. In addition other comparisons between human ratings and scene features have been conducted. Asteriou et al. [2] showed that participants' risk and criticality ratings correlated with the number of traffic objects in a scene detected by the YOLO algorithm, and previous works utilized even simpler features, such as visual contrast or the JPG compression rate [2, 39]. Collectively, these studies indicate a strong interplay between the factors of gaze and visual scanning behavior, situation awareness, and perceived risk. Another avenue of research regarding risk perception is its connection to trust [26]. Being able to quantify the perceived risk and trust of a driver-passenger would allow the realization of various user interface proposals for trust calibration that suggest personalized feedback [7, 54, 55] following explainable AI (xAI) principles.

In this context, xAI methods could be used not only to develop adaptive HMIs but fulfill its originally intended purpose of explaining the characteristics of a prediction model. In this paper, we utilize this approach: identifying the factors that contribute to the concept of perceived risk the most.

2.3 Generating Explainable Predictions with SHAP

xAI seeks to make ML models more transparent by providing a suite of techniques that enable human users to understand, appropriately trust, and produce more explainable models [12, 15, 27, 46]. The explanations xAI algorithms offer can be understood as an interface between humans and models [3], offering valuable insights across various applications like transportation [18] or healthcare [56]. In the context of risk perception, explainable predictions are especially important. While predicting perceived risk is useful, particularly for automated driving systems, our goal goes further. In this work, we aim to better understand the factors that influence how people perceive risk, with a focus on visual attention allocation. Which elements in a scene contribute most to risk prediction? And how does perceived risk vary across individuals?

To generate such explanations we leverage SHapley Additive exPlanation (SHAP). SHAP, proposed by Lundberg and Lee [31] is a method based on Shapley values from coalitional game theory [45]. The SHAP method provides local and global explanations for arbitrary predictive models. Various methods exist for approximating SHAP values across different types of ML models. In this study, we employ TreeSHAP [29], which enables the exact computation of SHAP values for tree-based models. Compared to methods like LIME [42] or tree-specific techniques, such as permutation importance and feature impurity calculations, SHAP offers several advantages. Grounded in game theory [35], SHAP values provide theoretical guarantees of consistency and local accuracy. Moreover, they ensure alignment between local and global explanations, clearly indicate whether each feature's contribution is positive or negative, and, as demonstrated by Lundberg et al., exhibit a stronger correspondence with human intuition [29, 31].

3 Methods

The following section outlines the user study conducted for data collection, the subsequent data processing steps, and the ML approach employed to model and explain drivers' risk perception based on SA and visual attention allocation.

3.1 User Study and Data Collection

In our user study, participants continuously rated the perceived risk while watching dashcam footage of inner-city driving. Throughout the experiment, their gaze behavior was recorded. The study was approved by the ethics committee of Leipzig University.

Participants. We recruited 27 participants through opportunity sampling at our institution (17 identified as male, 10 as female, and 1 as non-binary; Mean = 35.2 years, SD = 12.1). All participants held a valid driver's license. To ensure compatibility with the eye-tracking system, individuals who wore glasses were excluded from the study.

Apparatus. To represent typical inner-city traffic, we used a dashcam video (Full HD, 60 fps) recorded in Berlin obtained from YouTube. We obtained written permission from the creator to use the video in this study. To cover a wide range of risk ratings, we selected four segments with partially critical situations and four with predominantly non-critical situations. To cover the heterogeneity of urban traffic, we selected the segments to vary in traffic density, road type, and road user density. The participants viewed the videos through a VARJO XR-3 VR headset to enhance the immersive experience and allow eye-tracking. To enable real-time, continuous assessment of perceived risk, we developed a custom controller featuring a rotary knob as described in [2]. The device comprises an Arduino board connected to a potentiometer, all enclosed within a 3D-printed casing. A manual on how to build and 3D-print the controller is provided at Sup. *Controller Manual*. To show the videos and synchronize them with eye-tracking data and risk ratings, we developed a Unity application that displays the videos on a virtual canvas ($200cm \times 112,5cm$) in $130cm$ distance to the headset's camera view. Head movement and eye-tracking are synchronized using the Varjo Base software [1]. Communication between the controller and the Unity application is handled using the *Uduino* package [2]. The controller input (rotation angle) is linearly mapped to a risk score between 0 and 100. Building on the results of a preliminary pilot study, we implemented peripheral visual feedback using an ambient light display surrounding the video display. The ambient light continuously reflects the risk rating using a traffic light color scheme from green (low), over yellow (medium), to red (high). This allows for an intuitive interpretation without distracting from the driving scene or interfering with gaze behavior during the experiment.

Study Procedure. At the beginning, participants were given written and verbal information about the study procedures and potential VR-related risks, such as motion sickness. They were then assigned a pseudonym for data collection and completed a demographic questionnaire on gender, age, driving habits, and risk disposition. The demographic data is available at Sup. *Demographics*. Before the experiment started, we calibrated the eye-tracker and asked participants to familiarize themselves with the controller's haptics and the ambient light feedback. Subsequently, the participants viewed eight dashcam videos. Eight 35-second videos were presented in random order to reduce order effects. Between the videos we showed a 5-second countdown to allow participants to return the controller to the neutral position to ensure a neutral risk assessment at the beginning of each video. After the experiment, we conducted a semi-structured interview to gain additional insight into subjective perception, perceived gaze behavior, and risk assessment.

[1]https://varjo.com/use-center/get-to-know-your-headset/using-varjo-base/
[2]https://marcteyssier.com/uduino/

3.2 Data Processing

In this section we describe how we process the raw data to create a dataset on which a ML model can be trained. For all recorded data per 35 second trial we discard the first 5 seconds, because we consider this as the adaption phase where participants have to familiarize with the new scene before assessing perceived risk.

Video Data. To obtain scene features we apply YOLOv11 on all videos. For each frame we infer the object bounding boxes, including id, position, width and height, certainty of prediction, and the class of the object. The data is available at Sup. *YOLO BoundingBoxes*.

Gaze Data. To analyze participants' gaze behavior, we first match all gaze coordinates per frame to the bounding boxes detected by YOLO. To account for slight inaccuracies in bounding box detection and gaze behavior, we consider a gaze to be *"on"* an object if the gaze point is within 35 pixels of a YOLO bounding box. All gaze points that do not meet this criterion are assigned to a separate category. Subsequently, we compute the Shannon entropy of fixated objects per frame across all participants $H = -\sum_{i \in IDs} p_i \log_2 p_i$ over all object categories, where $p_i = \frac{|\text{gaze}_i|}{|IDs|}$.

Risk Ratings. Since the recorded risk data varies significantly between participants (see raw values as shown in Figure 2), we normalize the perceived risk values using min-max normalization over all videos per participant. Accordingly, for each participant, 100 represents the highest risk value and 0 represents the lowest risk value.

3.3 Data Aggregation

To capture the temporal dynamics of the driving scene in the context of continuous risk ratings, we employ a sliding window approach. Specifically, we aggregate the preprocessed data using the mean in overlapping time windows of 2 seconds and a step size of 0.5 seconds. A window length of 2 seconds can be considered an adequate proxy for capturing small changes and essential detail in the scene. This value is based on the guideline for off-road glances, as glances longer than 2 seconds are considered potentially dangerous in dynamic environments [24], where much can change within this short time span. To ensure sufficient differences between the chunks for the ML model, we use a step size of 0.5 seconds.

3.4 Modeling and Feature Selection

We use XGBoost as our modeling approach due to its strong predictive performance on tabular data, its simplicity, and explainability via TreeSHAP. For training we use mean squared error (MSE) as loss function, a learning rate of 0.01, and a maximum of 800 boosting iterations. To mitigate overfitting, early stopping is applied with a patience of 20 iterations. The final input vector used for prediction includes participant demographics as well as scene and gaze features. To test individual demographic features and their dependence on perceived risk, we include *age*, and *frequency of use*. The derived feature d_{age} represents the participant's age (age_p) in years, which is encoded into the following bins using numerical values for labeling: $age_p < 30$ years (0); $30 \leq age_p < 40$ years (1); $age_p \geq 40$ years (2). We encode the *frequency of use* d_{usage} into three bins as well, namely *at least once per week* (2), *at lest once a month but less than once a week* (1), and *less than once a month* (0).

For the detected YOLO features we only keep bounding boxes of the following categories *person, bicycle, car, motor cycle, bus, truck, traffic light, stop sign, dog*. Based on the bounding box data we calculate the maximal bounding box size and the sum of all bounding boxes per category visible in one frame. The resulting features are denoted as $s_{\text{num\_class}}$. Similar to related work De Winter et al. [10], we compute the size of each bounding box in pixels and introduce the feature $s_{max\_bb\_class}$, which represents the maximal bounding box size for each class in a frame, to approximate the distance between ego vehicle and object. For all $s_{\text{num\_class}}$ and $s_{\text{max\_bb\_class}}$ features, we additionally compute the change rates $s_{\text{slope\_num\_class}}$ and $s_{\text{slope\_max\_bb\_class}}$ per chunk to better capture temporal differences. In addition to the object detection, we carry out manual labeling to capture the context of the driving situation more precisely. We determine the number of lanes s_{lanes} (1 to 4) and proximity to a junction $s_{\text{proximity junction}}$ which is encoded numerically (*not in sight* (0); *in sight* (1); *on junction* (2)) for each video frame. For every YOLO class we determine how much time of a chunk a participant looked at this specific class. This results in a feature for every class, denoted as $g_{\text{class name}}$.

4 Evaluation and Results

In the following, we present the collected dataset, report on the performance of our modeling approach, and present the explainable predictions generated with SHAP.

4.1 Dataset

The final dataset comprises 388, 214 annotated video frames collected from 27 participants. For training and evaluating the ML models, we use a subset of 12, 312 samples, which were generated by segmenting the full dataset into temporal chunks, as described in subsection 3.4. The distribution of risk ratings before normalization, aggregated per participant across all videos, is shown in Figure 2. There is considerable variability between participants. For example, some (such as P6 and P20) consistently gave low ratings, with a median of zero, while others (such as P4 and P8) had median ratings closer to 50. The Inter Quartile Range (IQR) also varies between participants, as seen in the differing box lengths (e.g., P5 and P9), and the total range of values differs as well, indicated by the whisker lengths (e.g., P1 and P9).

Figure 3 shows the mean risk rating and its standard deviation as well as the focus entropy over one video sequence. In this scene, a vehicle is overtaking the ego-vehicle from the right side, which is against the local traffic regulations. During the time of the overtaking maneuver the mean risk rating increases significantly ($25.8\,\text{s} \leq t \leq 28.3\,\text{s}$) while the standard deviation decreases, indicating consensus in the risk rating between participants. At the same time, focus entropy decreases. In this context, low entropy suggests that participants allocated their attention on similar objects, whereas high entropy reflects a more diverse distribution of visual focus across different objects. This pattern is also evident in the annotated saliency maps, where participants focused almost exclusively on the overtaking car. In contrast, in a different scenario where the ego vehicle drives through a one-lane construction zone, we observe a focus entropy of approximately 0.6, which is one of the lowest values across all videos. The saliency maps reveal that

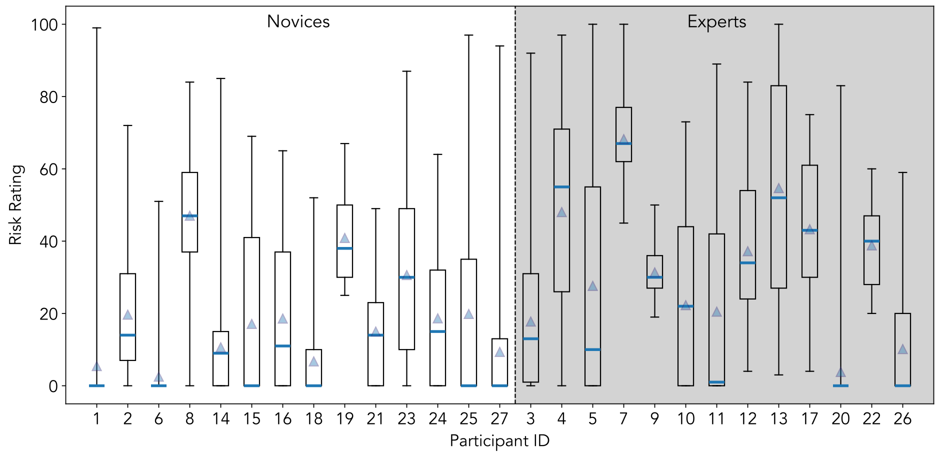


Figure 2: Risk ratings over all videos per participant before normalization clustered by experience (age, frequency of use).

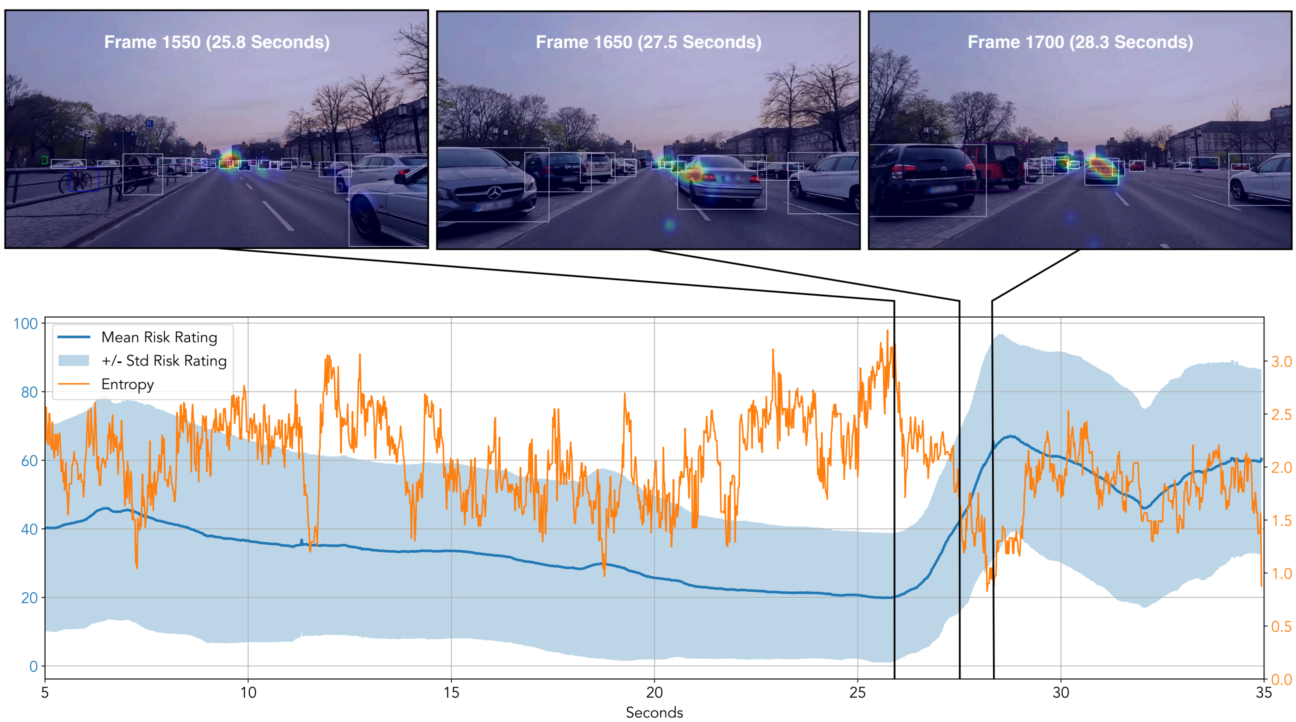


Figure 3: Visualization of how the mean risk rating, its standard deviation, and visual focus entropy vary over a video sequence. The annotated saliency maps visualize the participants' gaze patterns in three different situations, namely during a right-side overtaking maneuver, which led to an increase in risk rating and a decrease in entropy.

most participants visually fixated on the same object: the leading car (Sup. *HeatMap*). However, in this case, shared visual focus did not reduce the standard deviation of risk ratings. This suggests that even though participants looked at the same object, their perceived risk varied significantly.

4.2 Experimental Results

Apart from evaluating how accurately we can predict perceived risk in traffic situations, we are also interested in whether gaze information (i.e., where do drivers look at) improves prediction performance, and how well our ML models handle the high variability in participants' risk ratings (see Figure 2). To this end, we conduct

Table 1: Comparison of the different models. It shows the MAE on the participant level (leave-one-video-out cross validation), over all participants (leave-one-participant-out cross validation), and leave-one-participant-out over the novice and expert clusters derived by agglomerative clustering.

Model	Individual Participants		All Participants	Agglomerative Cluster	
	MAE	Std	MAE	MAE Novice	MAE Expert
XGBoost (without Gaze)	**19.631**	**7.056**	**19.406**	**17.046**	**22.039**
Mean Predictor	21.083	7.239	25.055	23.670	25.692
Linear Regression	33.454	9.139	23.424	18.994	23.170
XGBoost (with Gaze)	**19,914**	**7.158**	**19.771**	**17.554**	**22.271**
Mean Predictor	21.083	7.239	25.055	23.670	25.692
Linear Regression	34.638	9.466	23.429	19.214	23.319

three sets of experiments using different subsets of our dataset: individual participants, all participants combined, and participant clusters (see Table 1). In each experiment, we compare an XGBoost model to a linear regression model and a baseline mean predictor. Additionally, we assess the impact of gaze information by comparing models that incorporate gaze features to those that do not. Due to our relatively small dataset (N=27) and multiple samples per participant (2 second chunks that overlap), we evaluate all models using leave-one-out cross-validation.

Leave-One-Participant-Out Cross-Validation over all Participants. In the first experiment, we perform a leave one out evaluation across the participants. This means that we train a model over all participants except one. The data of this participant is used for evaluation. This procedure is applied to all participants, based on which we then compute the mean error over all left out participants. For the XGBoost model the average Mean Absolute Error (MAE) over all left out participants is $MAE = 19.406$ for the no-gaze experiment and $MAE = 19.771$, respectively. The mean predictor and linear regression models performed significantly worse (see Table 1).

Clustered by Experience. In the second experiment, we cluster participants using agglomerative clustering with respect to age and car usage frequency, both of which have been shown to influence risk perception [36]. The rationale behind this clustering is to investigate whether or not it is easier to predict perceived risk for one of the two groups. The two resulting clusters are of size $c_{\text{expert}} = 13$ and $c_{\text{novice}} = 14$, with c_{expert} containing the more experienced participants (see Figure 2). For training and evaluation we perform leave-one-out cross-validation only on the separate clusters, not on the whole dataset. The cluster c_{novice} achieves the most accurate prediction for gaze ($MAE = 17.554$) and no-gaze ($MAE = 17.046$) models over all experiments. However, the cluster c_{expert} representing the more-experienced drivers has the largest error over all experiments ($MAE_{gaze} = 22.271$, $MAE_{no-gaze} = 22.039$).

Individual Leave-One-Video-Out-Predictions. We test predictions on an individual level in the third experiment. Therefore, we perform a leave-one-out evaluation over the videos for each participant separately. For evaluation, we average the individual resulting losses over all participants. Compared to the other experiments, the MAE is close to the results for the leave-one-out evaluation over all participants for the gaze ($MAE = 19.914$) and no-gaze models ($MAE = 19.631$).

4.3 Explainable Predictions

Being able to predict perceived risk is valuable for in-car applications that for example use these predictions to trigger explanations [16, 53] or adapt their design. However, black box predictions are of limited value when it comes to understanding human behavior. To generate deeper insights into the key factors influencing perceived risk most, we use SHAP. SHAP enables us to identify the most relevant features contributing to model predictions (Figure 4) and to examine how these features interact, helping to explain their influence on the model's output. Despite the small difference in MAE between gaze and no-gaze models (compare Table 1), we apply SHAP to the XGBoost model trained with gaze features and across all participants. This allows us to analyze trends in the gaze features, even though their importance for the overall prediction is comparatively low.

To understand which features impact the model prediction most, we grouped the most predictive ones in a beeswarm plot shown in Figure 4. In a beeswarm plot each row represents one feature and each dot in a row corresponds to an individual prediction (i.e., one 2 second traffic scene) with the color representing the feature value. The features are ordered by the mean absolute value of the SHAP values with the most important features at the top. The position on the x-axis indicates the SHAP value, reflecting the impact this feature has on the perceived risk prediction. d_{age}, d_{usage}, and $s_{\text{max\_bb\_car}}$ are the most predictive features. Whereby higher age values contribute to a higher perceived risk prediction and medium age values contribute to a lower risk prediction. However, low age values have no direct impact on the model output in any direction. The plot also shows that the higher participants' usage frequency d_{usage} the lower is the predicted perceived risk value. Interestingly, scenes with a higher number of cars are associated with lower predicted perceived risk, as indicated by $s_{\text{num\_car}}$. In Figure 4 we also observe that a decrease in $s_{\text{max\_bb\_car}}$ is associated with lower perceived risk predictions. We use $s_{\text{max\_bb\_car}}$ as a proxy for the distance from the ego car, with smaller values indicating that the object is further away. The first gaze related feature g_{cars} is the 7th most predictive. There is no clear trend in how this feature affects model prediction. The sum of the remaining 39 features indicates that while some high feature values influence model predictions, the majority have an effect close to zero, consistent with these features being ranked less important.

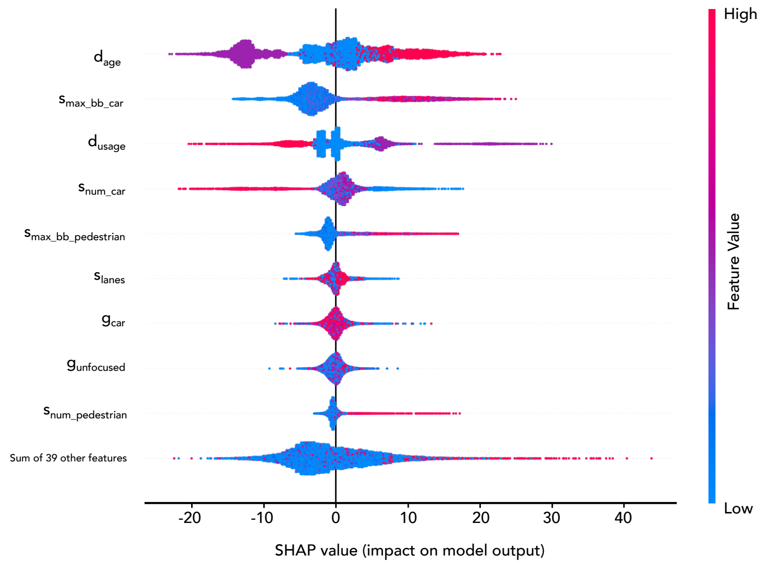


Figure 4: Beeswarm plots for the leave one out validation over all participants. Every row represents a feature. Each point in a row corresponds to an individual prediction and is colored according to the feature value.

To understand the effect of single feature values on the model's prediction in more detail, we plot the SHAP values (y-axis) against the corresponding feature values (x-axis). Every aggregated chunk (a 2 second driving scene) is represented as a dot (see Fig. 5). Vertical dispersion at a given x-value indicates non-linear dependencies between the displayed feature and other features. The colors represent the feature values of the feature with the highest interaction effect. The histogram at the bottom of each plot illustrates the distribution of data points across the feature range. Fig 5a shows that using a car *daily* (2) leads to lower risk predictions. For participants who use a car at most once a month or less, there is a neutral to slightly positive or negative effect on the model output, depending on age. Using a car weekly, but not daily, has the strongest effect on risk perception, especially for people aged between 30 and 40. Figure 5b indicates that for situations with a low number of cars ($s_{\text{num\_cars}} \leq 5$) risk predictions tend to be higher, whereas for a medium number of cars ($5 < s_{\text{num\_cars}} \leq 16$) there is no affect on risk predictions. However, a high number of cars ($s_{\text{num\_cars}} \geq 16$) leads to lower risk predictions, indicating that scenes with lots of cars are perceived as less risky. As shown in Fig 4, the most important gaze features by mean SHAP value are g_{car} and $g_{\text{unfocused}}$. However, they remain close to zero across their respective value ranges as seen in Figure 5d for g_{car}. This suggests that they have little effect on the prediction of perceived risk.

Fig 5c shows that smaller bounding box sizes up to 90, 000 pixels (4.3% of the screen size 1920×1080 pixels) have a negative impact on the model output. This indicates perceived risk is lower when other cars in the scene are far away, even if there are many cars. However, average maximal car bounding box areas per chunk larger than 90, 000 pixels (4.3% of screen size) lead to high positive model outputs, regardless of the number of cars in the scene, what indicates that proximity to other cars influences perceived risk.

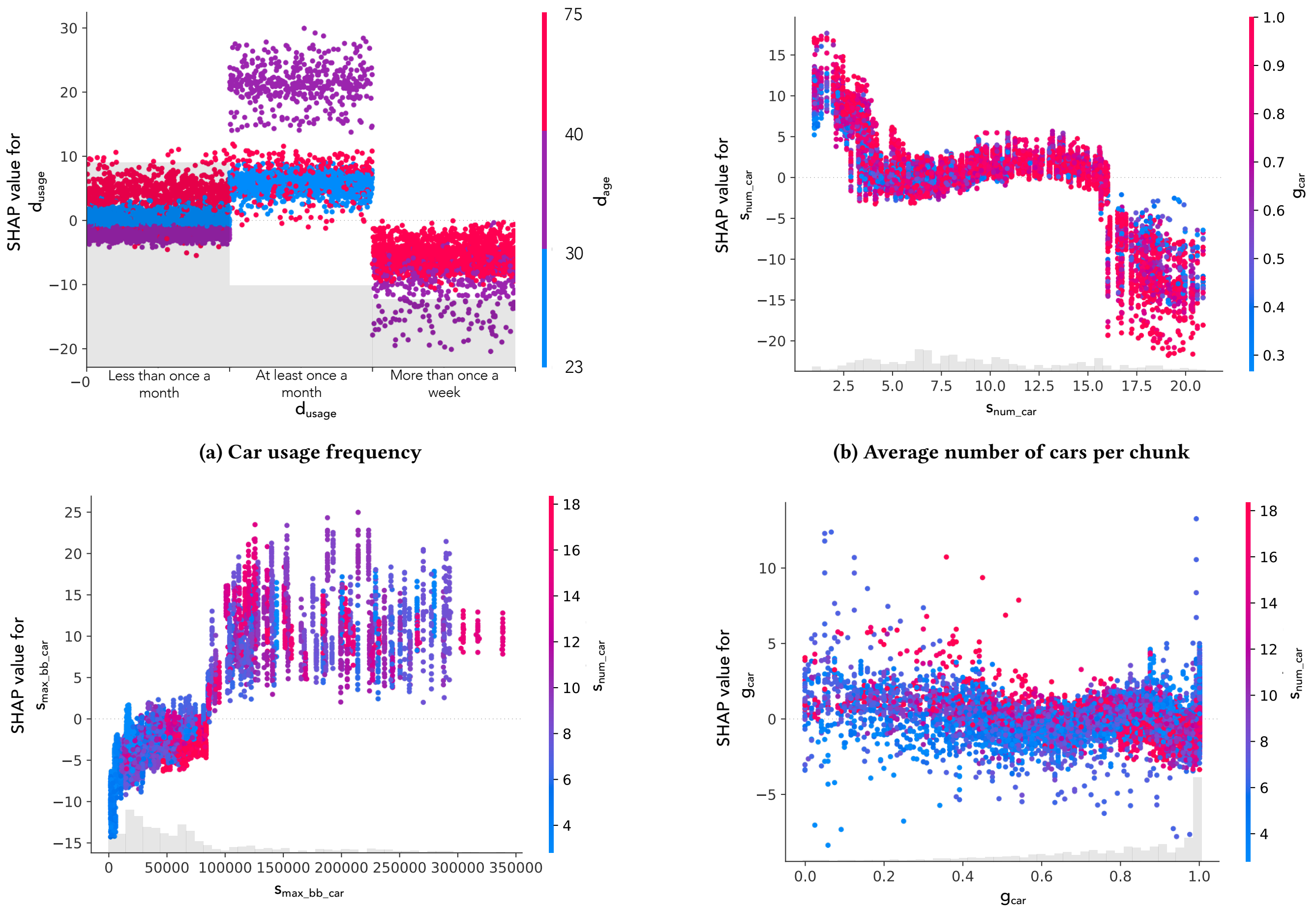
5 Discussion

Although the information where drivers look at did not enhance prediction accuracy, this outcome is itself an interesting and valuable finding. While the results may not show a significant boost in predictive performance, the combination of descriptive analysis and explainable predictions provides meaningful and insightful contributions to our understanding of the perceived risk in driving.

5.1 Risk Perception is Primarily Shaped by Individual Differences and the Driving Context

Our results show that individual characteristics, such as age and driving frequency, as well as contextual factors like the number of nearby vehicles and proximity to other cars or pedestrians, are among the most influential features for predicting perceived risk (see Figure 4). The finding that risk perception is shaped by age and driving experience aligns with prior research, which shows, for example, that drivers with different levels of experience respond differently to hazards [8, 9], and that older drivers tend to perceive driving situations as more risky [49]. However, as noted by Lorenz et al. [28], current modeling approaches often overlook such demographic information. These insights suggest a promising direction for future work: incorporating individual differences into model design may enable more personalized and accurate predictions of perceived risk.

Our approach further shows that our models did a better job in predicting perceived risk for novice drivers than for experienced drivers. For the latter group the prediction error was more than 5 points worse. This might be an indication that experienced drivers assess risk on a more nuanced level than novices do. Interestingly, an increase in the number of cars per scene is associated with a decrease in predicted risk, while a higher number of pedestrians is



(a) Car usage frequency

(b) Average number of cars per chunk

(c) Maximum average car bounding box size in pixels per chunk

(d) Fraction of time per chunk looking at a car

Figure 5: Feature dependency plots for selected features

correlated with an increase in perceived risk. One possible explanation is that scenes with many cars may serve as a proxy for larger, more structured streets, with clearer traffic rules, fewer driveways, and limited pedestrian crossings, and thus appear safer. A related finding was reported by De Winter et al. [10], who observed that higher vehicle speeds in their model were associated with lower perceived risk. They attribute this counterintuitive result to the concept of self-explaining roads which aims at delivering a road environment that aligns with drivers' expectations. Accordingly, roads with higher speed limits tend to be more structured and predictable, with clearer traffic patterns, which can lead to a lower perception of risk. Furthermore, our results indicate that the proximity to other road users, measured by the size of their bounding boxes, is a strong predictor of perceived risk. This finding is consistent with both computer vision-based approaches [10] and experimental studies [25]. In contrast to De Winter et al. [10], who used only static images, our study considers the slope of the bounding box size indicating if an object is approaching or moving away. However, this feature did not significantly impact model predictions.

5.2 Visual Sampling Behavior is a Weak Predictor of Perceived Risk in Driving

Our qualitative analysis (see subsection 4.1) and insights from explainable machine learning (see subsection 4.3) suggest that perceived risk is shaped more by scene characteristics and individual differences than by the specific objects people focus on. Although participants tend to fixate on similar objects in highly critical events (see Figure 3), this shared gaze behavior also appears in scenarios where risk assessments diverge (Sup. *HeatMap*), indicating that shared visual focus does not imply shared risk perception. This is reflected in our modeling results: adding gaze data did not improve prediction accuracy (Table 1), suggesting it is not a strong predictor. Global explanations from our XGBoost model reinforce this, showing that scene-related and demographic features contribute more to predictions than gaze-related ones. Only 2 of the top 10 predictive features involve gaze (Figure 4).

In our experiment, participants continuously scanned for risk cues while rating perceived risk in real time. In unanimously critical scenes, visual entropy was low (Figure 3), indicating that participants focused on the same key objects. This suggests such objects are perceived as high *value* as suggested in the SEEV model [52]. Yet, this did not translate into better model performance. Why? Because low visual entropy also occurs in scenes with widely varying risk ratings. In these cases, participants may expect relevant behavior from the object (high *expectancy*) but do not assign it much *value*, thus reducing its influence on perceived risk. Although our models take into account where participants direct their gaze, they do not capture why they look there or how they interpret what they see. Individual factors, such as age and driving experience, as well as latent personal factors that our model does not explicitly account for, play a larger role.

From an SA perspective [20, 21], our entropy findings suggest consistent *perception* (level 1 SA) for some situations, but high variability in *comprehension* and *projection* (SA levels 2 and 3), reinforcing our finding that risk perception is highly individual and experience-driven. Thus, while visual attention helps explain critical scenarios post-hoc, it is a weak standalone predictor of perceived risk in driving contexts.

5.3 Good and Bad News for Risk-Adaptive Automated Driving Technology

While automated vehicles are built to drive safely, they may overlook how safe the human driver feels. Predicting perceived risk using only the vehicle's sensor data could help tailor the driving experience, without the need for complex driver monitoring systems. In this regard, our results are good and bad news at the same time. On the one hand, the results suggest that incorporating fixation information is unlikely to significantly improve prediction accuracy, which can be considered encouraging, given that accurate and non-intrusive eye-tracking remains difficult to implement in vehicle cockpits. On the other hand, our modeling results highlight a key challenge: predicting perceived risk is inherently difficult, as individual risk perception varies widely. This variability makes it hard for manufacturers to create a one-size-fits-all model based solely on vehicle sensor data, such as inputs from cameras or radar. However, our cluster-based approach (compare Table 1) indicates that prediction accuracy can be improved for certain subgroups (novices in our case). Thus one potential solution could be to personalize risk prediction models by incorporating user-specific data obtained through connectivity services which are now commonly offered by most manufacturers [19]. These platforms could provide valuable contextual information about the driver's preferences, habits, or past behavior, enabling more adaptive and individualized safety systems. Furthermore, manufacturers have the advantage that they can compute the epistemic component of perceived risk such as the time to collision [22] which can be calculated based on measures retrieved from Advanced Driver Assistance Systems (ADAS).

5.4 Limitations and Future Work

Several limitations must be considered when interpreting the results of our study. First of all, our study is based on dashcam videos of traffic scenes shown on a large screen in VR, so participants could not utilize rearview or side mirrors or check the blind spots by turning their head. Future work could address this point by repeating the same experiment with 360° videos. Another limitation is that traffic is very diverse. We capture only a small amount of these situations. This shows in our leave-one-video-out predictions on individual participants, where the standard deviation in prediction is high. The training error is very low, but the evaluation error is very high for some videos, indicating low generalization capabilities. This is due to the fact that some situations in the test set are very different from what the model was trained on. Future work could address this issue by extending this dataset. To facilitate this process, we are making all of our artifacts available and hereby invite other researchers to contact us if they would like to conduct a similar study that would help expand this dataset.

Due to the small participant pool and the limited scenarios, our model could also be prone to overfitting. We took care to circumvent this by including as less as identifying participants' information as possible in the dataset and aggregated them into more general bins. However, a larger and more diverse participant pool would be valuable to validate our findings on a broader scale. Due to potential overfitting issues, the importance of demographic features must be partially attributed to latent individual differences, such as personal experiences, that our models do not account for, but which the explicit features we use serve as a proxy for.

As mentioned throughout the paper, we did not incorporate gaze-based metrics, such as fixation count or pupil diameter, into our models. While these features have been shown to predict risk or workload [5, 34], we were interested in where people look, what they focus on, and how these factors contribute to perceived risk. However, if the main goal is prediction accuracy, including such gaze-based metrics is a promising next step.

Open Science Statement. To improve openness and transparency in automotive user research [17], we make all research artifacts, including a description on how to use and extend this dataset, available on OSF: https://osf.io/cwd6h.

Acknowledgments

The authors acknowledge the financial support by the Federal Ministry of Research, Technology and Space of Germany and by Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus in the programme Center of Excellence for AI-research „Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig", project identification number: ScaDS.AI.

Further, this publication is supported by the Austrian Research Promotion Agency (FFG) under Grant No. 917805 (AttDrive).

Finally, we want to thank Oliver Welz for printing and manufacturing the case for the rotary knob, Celine Conzen and Timur Getselev for helping with the Unity implementation, and Cristina Bayer for her contributions to the pilot study.

References

[1] 2021. *SAEJ3016: Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles.* Standard. Society of Automotive Engineers (SAE), Warrendale. tex.key: SAEJ3016.

[2] Philipp Michael Markus Peter Asteriou, Heike Christiane Kotsios, and Philipp Wintersberger. 2024. What Characterizes "Situations" in Situation Awareness? Findings from a Human-centered Investigation. In *Proceedings of the 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, Stanford CA USA, 216–226. doi:10.1145/3640792.3675722

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (June 2020), 82–115. doi:10.1016/j.inffus.2019.12.012

[4] P. Bazilinskyy, Y. B. Eisma, D. Dodou, and J. C. F. De Winter. 2020. Risk perception: A study using dashcam videos and participants from different world regions. *Traffic Injury Prevention* 21, 6 (Aug. 2020), 347–353. doi:10.1080/15389588.2020.1762871

[5] Olga Vl. Bitkina, Jaehyun Park, and Hyun K. Kim. 2021. The ability of eye-tracking metrics to classify and predict the perceived driving workload. *International Journal of Industrial Ergonomics* 86 (Nov. 2021), 103193. doi:10.1016/j.ergon.2021.103193 Publisher: Elsevier BV.

[6] Matthew L Bolton, Elliot Biltekoff, and Laura Humphrey. 2021. The level of measurement of subjective situation awareness and its dimensions in the situation awareness rating technique (SART). *IEEE Transactions on Human-Machine Systems* 52, 6 (2021), 1147–1154.

[7] Mark Colley, Max Rädler, Jonas Glimmann, and Enrico Rukzio. 2022. Effects of Scene Detection, Scene Prediction, and Maneuver Planning Visualizations on Trust, Situation Awareness, and Cognitive Load in Highly Automated Vehicles. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (July 2022), 1–21. doi:10.1145/3534609

[8] David Crundall. 2016. Hazard prediction discriminates between novice and experienced drivers. *Accident Analysis & Prevention* 86 (Jan. 2016), 47–58. doi:10.1016/j.aap.2015.10.006

[9] David Crundall, Peter Chapman, Steven Trawley, Lyn Collins, Editha Van Loon, Ben Andrews, and Geoffrey Underwood. 2012. Some hazards are more attractive than others: Drivers of varying experience respond differently to different types of hazard. *Accident Analysis & Prevention* 45 (March 2012), 600–609. doi:10.1016/j.aap.2011.09.049

[10] Joost De Winter, Jim Hoogmoed, Jork Stapel, Dimitra Dodou, and Pavlo Bazilinskyy. 2023. Predicting perceived risk of traffic scenes using computer vision. *Transportation Research Part F: Traffic Psychology and Behaviour* 93 (Feb. 2023), 235–247. doi:10.1016/j.trf.2023.01.014

[11] Hanwen Deng, Guoliang Xiang, Jiandong Pan, Xianhui Wu, Chaojie Fan, Kui Wang, and Yong Peng. 2024. How to design driver takeover request in real-world scenarios: A systematic review. *Transportation Research Part F: Traffic Psychology and Behaviour* 104 (July 2024), 411–432. doi:10.1016/j.trf.2024.06.012

[12] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[13] Tom Driessen, Dimitra Dodou, Pavlo Bazilinskyy, and Joost De Winter. 2024. Putting ChatGPT vision (GPT-4V) to the test: risk perception in traffic images. *Royal Society Open Science* 11, 5 (May 2024), 231676. doi:10.1098/rsos.231676

[14] Na Du, Xingwei Wu, Teruhisa Misu, and Kumar Akash. 2022. A preliminary study of modeling driver situational awareness based on SEEV and ACT-R Models. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 66. SAGE Publications Sage CA: Los Angeles, CA, 1602–1606.

[15] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, and Rajiv Ranjan. 2023. Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *Comput. Surveys* 55, 9 (Sept. 2023), 1–33. doi:10.1145/3561048

[16] Patrick Ebel. 2024. Generative AI and Attentive User Interfaces: Five Strategies to Enhance Take-Over Quality in Automated Driving. doi:10.48550/ARXIV.2402.10664 Version Number: 1.

[17] Patrick Ebel, Pavlo Bazilinskyy, Mark Colley, Courtney Michael Goodridge, Philipp Hock, Christian P. Janssen, Hauke Sandhaus, Aravinda Ramakrishnan Srinivasan, and Philipp Wintersberger. 2024. Changing Lanes Toward Open Science: Openness and Transparency in Automotive User Research. In *Proceedings of the 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, Stanford CA USA, 94–105. doi:10.1145/3640792.3675730

[18] Patrick Ebel, Christoph Lingenfelder, and Andreas Vogelsang. 2023. On the Forces of Driver Distraction: Explainable Predictions for the Visual Demand of in-Vehicle Touchscreen Interactions. *Accident Analysis & Prevention* 183 (April 2023), 106956. doi:10.1016/j.aap.2023.106956

[19] Patrick Ebel, Julia Orlovska, Sebastian Hünemeyer, Casper Wickman, Andreas Vogelsang, and Rikard Söderberg. 2021. Automotive UX design and data-driven development: Narrowing the gap to support practitioners. *Transportation Research Interdisciplinary Perspectives* 11 (Sept. 2021), 100455. doi:10.1016/j.trip.2021.100455 Publisher: Elsevier BV.

[20] Mica R. Endsley. 1995. Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 37, 1 (March 1995), 32–64. doi:10.1518/001872095779049543

[21] Mica R. Endsley. 2021. A Systematic Review and Meta-Analysis of Direct Objective Measures of Situation Awareness: A Comparison of SAGAT and SPAM. *Human Factors* 63, 1 (2021), 124–150. doi:10.1177/0018720819875376

[22] Sven Ove Hansson. 2010. Risk: objective or subjective, facts or values. *Journal of Risk Research* 13, 2 (March 2010), 231–238. doi:10.1080/13669870903126226

[23] Rebecca Ivers, Teresa Senserrick, Soufiane Boufous, Mark Stevenson, Huei-Yang Chen, Mark Woodward, and Robyn Norton. 2009. Novice drivers' risky driving behavior, risk perception, and crash risk: findings from the DRIVE study. *American journal of public health* 99, 9 (2009), 1638–1644.

[24] Sheila Klauer, Thomas Dingus, T Neale, J. Sudweeks, and D Ramsey. 2006. *The Impact of Driver Inattention on Near-Crash/Crash Risk: An Analysis Using the 100-Car Naturalistic Driving Study Data*. Technical Report. U.S. Department of Transportation, National Highway Traffic Safety Administration / Virginia Tech Transportation Institute, 3500 Transportation Research Plaza (0536) Blacksburg, Virginia 24061. Volume: 594.

[25] Takayuki Kondoh, Tomohiro Yamamura, Satoshi Kitazaki, Nobuyuki Kuge, and Erwin Roeland Boer. 2008. Identification of Visual Cues and Quantification of Drivers' Perception of Proximity Risk to the Lead Vehicle in Car-Following Situations. *Journal of Mechanical Systems for Transportation and Logistics* 1, 2 (2008), 170–180. doi:10.1299/jmtl.1.170

[26] Mengyao Li, Brittany E Holthausen, Rachel E Stuck, and Bruce N Walker. 2019. No risk no trust: Investigating perceived risk in highly automated driving. In *Proceedings of the 11th international conference on automotive user interfaces and interactive vehicular applications*. 177–185.

[27] Zachary C. Lipton. 2018. The Mythos of Model Interpretability. *Queue* 16, 3 (June 2018), 31–57. doi:10.1145/3236386.3241340 Publisher: Association for Computing Machinery (ACM).

[28] Martin Lorenz, Tiago Amorim, Debargha Dey, Mersedeh Sadeghi, and Patrick Ebel. 2024. Computational Models for In-Vehicle User Interface Design: A Systematic Literature Review. In *Proceedings of the 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, Stanford CA USA, 204–215. doi:10.1145/3640792.3675735

[29] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2, 1 (Jan. 2020), 56–67. doi:10.1038/s42256-019-0138-9 Publisher: Springer Science and Business Media LLC.

[30] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. 2018. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888* (2018).

[31] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4768–4777. event-place: Long Beach, California, USA.

[32] Andrew K. Mackenzie and Julie M. Harris. 2015. Eye movements and hazard perception in active and passive driving. *Visual Cognition* 23, 6 (July 2015), 736–757. doi:10.1080/13506285.2015.1079583 Publisher: Informa UK Limited.

[33] Sarah Malone and Roland Brünken. 2020. Studying Gaze Behavior to Compare Three Different Hazard Perception Tasks. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 62, 8 (Dec. 2020), 1286–1303. doi:10.1177/0018720819873462 Publisher: SAGE Publications.

[34] Gerhard Marquart, Christopher Cabrall, and Joost De Winter. 2015. Review of Eye-related Measures of Drivers' Mental Workload. *Procedia Manufacturing* 3 (2015), 2854–2861. doi:10.1016/j.promfg.2015.07.783 Publisher: Elsevier BV.

[35] Christoph Molnar. 2020. *Interpretable machine learning*. Lulu.com. https://www.ebook.de/de/product/39178906/christoph_molnar_interpretable_machine_learning.html tex.ean: 9780244768522.

[36] Caroline Moran, Joanne M. Bennett, and Prasannah Prabhakharan. 2019. Road user hazard perception tests: A systematic review of current methodologies. *Accident Analysis & Prevention* 129 (Aug. 2019), 309–333. doi:10.1016/j.aap.2019.05.021

[37] John Han Numan. 1998. Knowledge-based systems as companions: Trust, human computer interaction and complex systems. (1998).

[38] Daniel Omeiza, Helena Webb, Marina Jirotka, and Lars Kunze. 2022. Explanations in Autonomous Driving: A Survey. *IEEE Transactions on Intelligent Transportation Systems* 23, 8 (Aug. 2022), 10142–10162. doi:10.1109/TITS.2021.3122865

[39] Sami Park, Yilun Xing, Kumar Akash, Teruhisa Misu, and Linda Ng Boyle. 2022. The Impact of Environmental Complexity on Drivers' Situation Awareness. In *Proceedings of the 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, Seoul Republic of Korea, 131–138. doi:10.1145/3543174.3546831

[40] Anuj Kumar Pradhan, Kim R. Hammel, Rosa DeRamus, Alexander Pollatsek, David A. Noyce, and Donald L. Fisher. 2005. Using Eye Movements To Evaluate Effects of Driver Age on Risk Perception in a Driving Simulator. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 47, 4 (Dec. 2005), 840–852. doi:10.1518/001872005775570961

[41] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2015. You Only Look Once: Unified, Real-Time Object Detection. doi:10.48550/ARXIV.1506.02640 Version Number: 5.

[42] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. (2016). doi:10.48550/ARXIV.1602.04938 Publisher: arXiv Version Number: 3.

[43] Patrick Rossner and Angelika C. Bullinger. 2020. How Do You Want to be Driven? Investigation of Different Highly-Automated Driving Styles on a Highway Scenario. In *Advances in Human Factors of Transportation*, Neville Stanton (Ed.). Vol. 964. Springer International Publishing, Cham, 36–43. doi:10.1007/978-3-030-20503-4_4 Series Title: Advances in Intelligent Systems and Computing.

[44] Marlene Susanne Lisa Scharfe-Scherf, Sebastian Wiese, and Nele Russwinkel. 2022. A cognitive model to anticipate variations of situation awareness and attention for the takeover in highly automated driving. *Information* 13, 9 (2022), 418.

[45] L. S. Shapley. 1953. A Value for n-Person Games. In *Contributions to the Theory of Games (AM-28), Volume II*. Princeton University Press, 307–318. doi:10.1515/9781400881970-018

[46] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies* 146 (Feb. 2021), 102551. doi:10.1016/j.ijhcs.2020.102551

[47] Paul Slovic. 1987. Perception of risk. *science* 236, 4799 (1987), 280–285.

[48] Rachel E Stuck, Brianna J Tomlinson, and Bruce N Walker. 2022. The importance of incorporating risk into human-automation trust. *Theoretical Issues in Ergonomics Science* 23, 4 (2022), 500–516.

[49] Ulrich Tränkle, Christhard Gelau, and Thomas Metker. 1990. Risk perception and age-specific accidents of young drivers. *Accident Analysis & Prevention* 22, 2 (April 1990), 119–125. doi:10.1016/0001-4575(90)90063-Q

[50] Pål Ulleberg and Torbjørn Rundmo. 2003. Personality, attitudes and risk perception as predictors of risky driving behaviour among young drivers. *Safety science* 41, 5 (2003), 427–443.

[51] Trent W. Victor, Emma Tivesten, Pär Gustavsson, Joel Johansson, Fredrik Sangberg, and Mikael Ljung Aust. 2018. Automation Expectation Mismatch: Incorrect Prediction Despite Eyes on Threat and Hands on Wheel. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 60, 8 (Dec. 2018), 1095–1116. doi:10.1177/0018720818788164

[52] Christopher D Wickens, Jason S McCarley, Amy L Alexander, Lisa C Thomas, Michael Ambinder, and Sam Zheng. 2005. Attention-Situation Awareness (A-SA) Model of Pilot Error. *Human Performance Modeling in Aviation* (2005).

[53] Gesa Wiegand, Malin Eiband, Maximilian Haubelt, and Heinrich Hussmann. 2020. "I'd like an Explanation for That!"Exploring Reactions to Unexpected Autonomous Driving. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, Oldenburg Germany, 1–11. doi:10.1145/3379503.3403554

[54] Philipp Wintersberger, Anna-Katharina Frison, Andreas Riener, and Tamara Von Sawitzky. 2018. Fostering User Acceptance and Trust in Fully Automated Vehicles: Evaluating the Potential of Augmented Reality. *PRESENCE: Virtual and Augmented Reality* 27, 1 (Feb. 2018), 46–62. doi:10.1162/pres_a_00320 Publisher: MIT Press.

[55] Philipp Wintersberger, Hannah Nicklas, Thomas Martlbauer, Stephan Hammer, and Andreas Riener. 2020. Explainable Automation: Personalized and Adaptive UIs to Foster Trust and Understanding of Driving Automation Systems. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, Virtual Event DC USA, 252–261. doi:10.1145/3409120.3410659

[56] Christopher C. Yang. 2022. Explainable Artificial Intelligence for Predictive Modeling in Healthcare. *Journal of Healthcare Informatics Research* 6, 2 (June 2022), 228–239. doi:10.1007/s41666-022-00114-1

[57] Yanqun Yang, Meifeng Chen, Changxu Wu, Said M Easa, and Xinyi Zheng. 2020. Structural equation modeling of drivers' situation awareness considering road and driver factors. *Frontiers in psychology* 11 (2020), 1601.

[58] Ting Zhang, Jing Yang, Nade Liang, Brandon J Pitts, Kwaku Prakah-Asante, Reates Curry, Bradley Duerstock, Juan P Wachs, and Denny Yu. 2023. Physiological measurements of situation awareness: a systematic review. *Human factors* 65, 5 (2023), 737–758.