OptiCarVis: Improving Automated Vehicle Functionality Visualizations Using Bayesian Optimization to Enhance User Experience

Pascal Jansen*
pascal.jansen@uni-ulm.de
Institute of Media Informatics, Ulm University
Ulm, Germany

Mark Colley*
mark.colley@uni-ulm.de
Institute of Media Informatics, Ulm University
Ulm, Germany
Cornell Tech
New York, U.S.

Svenja Krauß
svenja.krauss@uni-ulm.de
Institute of Media Informatics, Ulm University
Ulm, Germany

Daniel Hirschle
daniel.hirschle@uni-ulm.de
Institute of Media Informatics, Ulm University
Ulm, Germany

Enrico Rukzio
enrico.rukzio@uni-ulm.de
Institute of Media Informatics, Ulm University
Ulm, Germany

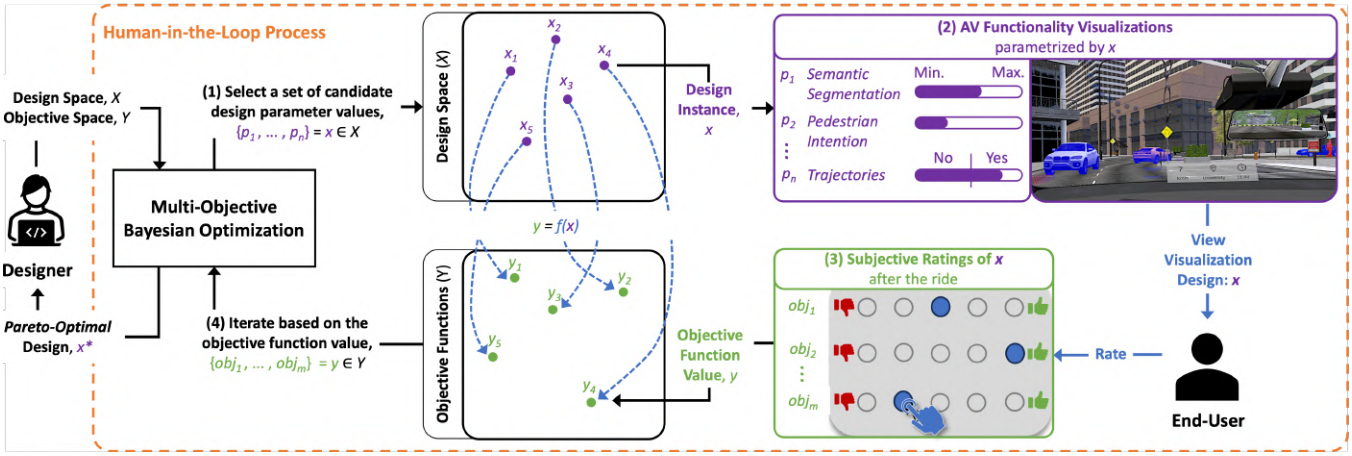


Figure 1: OptiCarVis—human-in-the-loop multi-objective Bayesian optimization of automated vehicle (AV) functionality visualization design to increase end-users' subjective ratings of design objectives, for example, trust, perceived safety, acceptance, and aesthetics, while reducing the cognitive load (obj_1 to obj_m). (1) OptiCarVis selects a set of parameter values (e.g., the color of semantic segmentation p_1 and whether to visualize vehicle trajectories p_n) from the design space X. (2) The end-user views the set of parameters x in a simulated AV ride and (3) returns subjective ratings. (4) In the next iteration, these are used as values y of the objective functions $f : X \rightarrow Y$ for which the design is optimized. Our approach finds a *Pareto-optimal* [61] visualization design x^* per end-user.

Abstract

Automated vehicle (AV) acceptance relies on their understanding via feedback. While visualizations aim to enhance user understanding of AV's detection, prediction, and planning functionalities, establishing an optimal design is challenging. Traditional "one-size-fits-all" designs might be unsuitable, stemming from resource-intensive empirical evaluations. This paper introduces OptiCarVis, a set of Human-in-the-Loop (HITL) approaches using Multi-Objective Bayesian Optimization (MOBO) to optimize AV feedback visualizations. We compare conditions using eight expert and user-customized designs for a *Warm-Start* HITL MOBO. An online study

*Both authors contributed equally to this research.

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713514

(N=117) demonstrates OptiCarVis's efficacy in significantly improving trust, acceptance, perceived safety, and predictability without increasing cognitive load. OptiCarVis facilitates a comprehensive design space exploration, enhancing in-vehicle interfaces for optimal passenger experiences and broader applicability.

CCS Concepts

• **Human-centered computing** → **Systems and tools for interaction design**; *Empirical studies in visualization*; **Empirical studies in HCI**.

Keywords

automated vehicles, user study, bayesian optimization, multi objective

ACM Reference Format:
Pascal Jansen, Mark Colley, Svenja Krauß, Daniel Hirschle, and Enrico Rukzio. 2025. OptiCarVis: Improving Automated Vehicle Functionality Visualizations Using Bayesian Optimization to Enhance User Experience. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3706598.3713514

1 Introduction

Driving automation is anticipated to alter mobility and traffic systems [30] fundamentally. According to the Society of Automotive Engineers (SAE) taxonomy J3016 [72], Automated Vehicles (AVs) range from Level 4 (conditional automation) to Level 5 (full automation). As users can engage in non-driving related activities while driving tasks are automated [44], design priorities have extended from mainly safety concerns to user experience. According to ISO 9241-210 [41], user experience incorporates all the users' emotions, beliefs, preferences, and perceptions before, during, and after use. In the AV context, user acceptance—the extent to which users are willing to use a new technology based on perceived ease of use and usefulness [23]—contributes to high user experience [14, 20, 73].

However, user acceptance of AV technology is not guaranteed. Studies have shown that many potential users are concerned about AV reliability [53, 74], which refers to AVs' ability to perform driving tasks safely across different situations consistently. Both undertrust (e.g., leading to not using AVs) and overtrust (e.g., inadequately supervising the operation) present challenges for AV usage if the user's trust is inappropriate to the actual AV reliability [55]. Therefore, previous works have shown that visualizations of AV functionality are a way to enhance user experience [20, 32, 73]. Suggested visualization solutions to overcome undertrust highlighted other road users in foggy scenarios [84]. In the Connected Automated Driving (CAD) context, proposals include visualizing external sensor coverage (henceforth called "CAD-covered area") and detected road users occluded, for example, by a building (henceforth called "occluded cars") [64]. Regarding overtrust, visualizing the internal functionalities of AVs (e.g., Situation Detection, Situation Prediction, or Trajectory Planning) and their inherent uncertainty were evaluated [13, 14, 20, 32].

However, the diversity of passengers complicates the design of AV functionality visualizations. Passengers' subjective perceptions of safety, trust, and aesthetics differ [60]. In addition, passengers' understanding of AVs' internal functionality depends on individual knowledge and attitudes towards technology [89]. Thus, to support user understanding, designers are tasked with balancing various design objectives within a complex design space (i.e., the set of possible design parameter values), for example, determining the size, transparency, and necessity of visualization elements. Traditional design methods relied on the user-centered design process (see ISO 9241 [41]), standards (e.g., ISO 15005 [39]), and guidelines (e.g., the JAMA Guidelines for In-vehicle Display System [87]) but also experience, trial and error, and intuition especially when designing for novel experiences as with AVs, resulting in resource-intensive user evaluations.

To align visualization designs with passengers' diverse needs and preferences, previous work explored personalization approaches. For instance, Normark [65] enabled passengers to manually personalize icons' size, location, and color on the dashboard, center stack, and Head-Up Display (HUD). However, without design experience, it can be challenging to translate personal needs and preferences into design parameter values. In contrast, computational methods have been used to help designers choose a visualization design. For example, Zhong et al. [91] elicited weights to designers' ratings, which informed a decision model to choose a design out of three candidate HUD design schemes. Likewise, Yunuo et al. [88] obtained weighted ratings by drivers and, accordingly, their model selected a HUD design out of 18 candidates. However, these previous works did not explore the design space with continuous values but only used a small set of pre-defined visualizations (see [88, 91]), which could potentially miss designs that better meet passengers' needs and preferences. Besides, while these works measure the usability [88, 91], they disregard perceived safety and trust, which are crucial for user acceptance of AVs [1] and, ultimately, the user experience.

If a design should enhance user experience, involving humans in the design process is important to ensure alignment with their needs and preferences. Therefore, a Human-in-the-Loop (HITL) process (see [9]) could be used that iteratively presents users with design variants. Their feedback is used in each iteration to optimize the design parameter values to meet the users' needs and preferences better. However, such an optimization is challenging because of multiple design objectives (e.g., increasing perceived safety and trust). In non-automotive User Interface (UI) domains, approaches using (Multi-Objective) Bayesian Optimization (BO and MOBO) have been emerging to solve design optimization problems [9, 10, 45, 49, 57]. BO optimizes designs by predicting which changes in the parameter values will most effectively meet the design objective. In MOBO–an extension of BO that can handle multiple objectives–trade-offs among potentially conflicting objectives are managed by identifying points on the *Pareto front*, where any improvement in one objective would lead to another objective's deterioration. For instance, on this front, enhancing perceived safety by showing more visualizations may increase cognitive load. Hereafter, one such design is referred to as **Pareto-optimal** [61].

HITL MOBO can explore large design spaces in just a few iterations. However, their effectiveness in optimizing visualizations of AV functionalities is uncertain, as this is determined by subjective passenger (hereafter referred to as "end-user") ratings such as perceived safety and trust, which can pose a challenge for optimization

(see [10]) compared to objective measures like input error rate or accuracy (see [45]). Besides, shortcomings are, for example, dealing with inconsistent human judgments [67] and disregarding users' prior knowledge and preferences, which may reduce agency and expressiveness [9, 57]. While previous works addressed this by involving designers in HITL MOBO [9, 49, 57], the potential effects of including end-users without technical or design backgrounds remain largely unexplored.

To overcome these limitations, we present OptiCarVis—the computational optimization of AV functionality visualization design using HITL MOBO. The design objectives are to increase end-users' perceived safety and trust in AV functionalities, their understanding of AVs' internal operations, and their perceived usefulness, satisfaction, and aesthetics of the visualizations while reducing their cognitive load. We involve end-users without technical or design backgrounds in HITL optimization to leverage their valuable knowledge, experiences, and preferences.

We demonstrate OptiCarVis by designing visualizations of an AV's functional levels (*Situation Detection*, *Situation Prediction*, and *Trajectory Planning*), as well as the CAD-covered area and occluded cars, and the general information (i.e., AV speed, AV destination, and current time) on a HUD, as these can convey relevant information on various stages of the automated driving task [20, 32, 64]. The visualizations include semantic segmentation for Situation Detection, pedestrian intention icons and other road users' trajectories for Situation Prediction, and a trajectory for the AV's own Trajectory Planning. We created an AV driving environment in Unity, where end-users view these visualizations on a simulated Augmented Reality (AR) Windshield Display (WSD) shown on their computer screen. In each HITL iteration (see Figure 1), the MOBO first estimates parameter values in the design space of functionality visualizations. End-users then experience the chosen values and provide subjective ratings **after** the ride. Next, the MOBO uses these ratings to understand how well the design has met the objectives and optimizes the parameter values for the subsequent iteration. This is repeated until design objectives are met (e.g., maximum trust on the rating scale).

Our work examines six design and optimization conditions: We use **(1)** No Visualization as a baseline. Additionally, to evaluate OptiCarVis against traditional design approaches, we include **(2)** a design created by automotive UI experts (N=8, using mean parameter values) and **(3)** custom designs by end-users. For our optimization conditions, we draw from previous works showing that MOBO initialized with prior data (i.e., knowledge about which design space area already achieves desired objectives; henceforth called *Warm-Start*) requires fewer iterations [58, 69]. Accordingly, besides **(4)** *Cold-Start* HITL MOBO with random initial design parameter values, **(5)** we employ a *Warm-Start* HITL MOBO initialized with mean parameter values from designs created by automotive UI experts (N=8). Lastly, **(6)** end-users create a custom design for their *Warm-Start* HITL MOBO.

To quantify the optimization's effectiveness, end-users evaluate their final design after the condition (regarding the cognitive load, predictability, trust, perceived safety, usefulness, satisfaction, and aesthetics). We also compare the HITL MOBO results to the baseline crafted by automotive UI experts (N=8). Besides, we capture end-user behaviors through webcam-based eye-tracking to monitor their reactions to MOBO-driven visualization designs and verify their attention throughout the study.

We conducted a between-subject online study with 117 participants. The study found that OptiCarVis returned personalized design parameters for visualizing AVs' Situation Detection, Situation Prediction, and own Trajectory Planning that significantly enhanced perceived safety relative to participants' custom design, improved predictability compared to the pre-defined expert design, and increased trust, usefulness, and satisfaction over both. Participants' feedback expressed satisfaction with the visualization design process, indicating a sense of involvement. Many deemed the design optimal but suggested incorporating additional visualization elements and expanding driving scenarios.

By leveraging OptiCarVis, automotive UI designers and end-users can navigate complex design spaces, potentially resulting in more passenger-centric UIs that could significantly increase their perceived safety, trust, and acceptance of AVs. Moreover, the direct integration of end-users' feedback into the design could inspire the development of more cooperative and effective end-user optimization methods that are implicitly integrated into future automotive UIs.

Contribution Statement: **(1)** OptiCarVis—the computational optimization of AV functionality visualization design using HITL MOBO to improve end-user trust, perceived safety, acceptance, and understanding of AVs and decrease cognitive load. **(2)** Empirical insights from a between-subject online study (N=117) assessing OptiCarVis against an averaged design by experts (N=8), end-users' custom designs, and a No Visualization baseline. The study also investigated how participants interacted with OptiCarVis in three optimization conditions: a Cold-Start HITL MOBO with random initial design parameters, a Warm-Start HITL MOBO initialized by mean parameters from expert designs (N=8), and a Warm-Start HITL MOBO initialized by end-users custom parameter values. **(3)** Open-source[1] implementation of a Unity-based driving simulation enabling HITL MOBO of AV functionality visualizations.

2 Background and Related Work

Our work builds on previous approaches in (1) visualizing AV functionalities, (2) personalization and computational methods for designing in-vehicle UIs, as well as (3) employing HITL MOBO.

2.1 In-Vehicle Visualizations of Automated Vehicle Functionalities

Previous research evaluated different display technologies (e.g., HUDs, LED strips, and AR WSDs) for visualizing various driving-related information in AVs. For instance, Colley et al. [13] found that the AR WSD reduced cognitive load. Besides, Häuslschmid et al. [36] increased trust by showing the AV's current situation interpretation of the vehicle through a miniature world or a simulated chauffeur avatar. They found the miniature world most effective, but participants' opinions on the necessity of such visualization varied significantly. Currano et al. [22] also tested an AR HUD and found situation awareness varied based on the driving scene's complexity and the participants' reported driving styles, suggesting that the HUD should be personalized and adapt to these factors.

[1] Source code will be released upon acceptance

Schneider et al. [73] evaluated explanations delivered via an AR WSD and an LED strip concerning the future vehicle trajectory. User experience increased with explanations, but adding a post-explanation via a smartphone app did not enhance it. Colley et al. [17] also found that an abstract visualization of objects (i.e., via a symbol with text) perceived by the AV, such as crossing dogs and children in a HUD, was sufficient.

An important factor contributing to understanding AV functionality and its safe use can be the visualization of uncertainties in detecting or predicting the driving environment and ego trajectory planning. Beller et al. [5] focused on conveying automation uncertainty with a simple anthropomorphic symbol when system limits are reached. They observed that situational awareness and trust increased when uncertainty information was displayed. Similarly, Helldin et al. [38] used abstract uncertainty visualization—bars indicating the system's ability to operate—and found that users took control more quickly. However, they also found that participants trusted automation less when uncertainty was shown. As AV uncertainties on the in-vehicle dashboard may increase distractions, unnecessary glances, and cognitive load, other visualization technologies were employed. Kunze et al. [51] used AR to visualize longitudinal and lateral control uncertainties. They found that the hue is particularly effective at conveying urgency.

However, these abstract visualizations (e.g., icons [5], ambient light [73], or miniature worlds [36]) may not allow the user to identify the source of uncertainty. Therefore, using semantic segmentation, Colley et al. [14] visualized AVs' Situation Detection. They found that their simulated AR WSD did not increase trust or cognitive load but improved situation awareness, and users rated detection-related attributes significantly better. In this context, Colley et al. [20] compared visualizations of the functional levels of AVs (detection, prediction, and trajectory planning) and their combinations. They found that showing the planned ego trajectory increased trust, and the prediction increased cognitive load. Flohr et al. [32] confirmed the need to visualize AV functionalities to combat over- and undertrust in AVs by visualizing Situation Detection in an on-road wizard-of-Oz study. They found that AV functionality visualizations can increase predictability, perceived usefulness, and hedonic user experiences. Lastly, Müller et al. [64] reused visualization designs, such as the connectivity symbol, planned trajectory, and vehicle markers. They also visualized infrastructure support via CAD and, enabled by this, displayed cars occluded behind obstacles and a gap indicator for AV merging. They found that combining all visualizations resulted in the highest trust, reliability, and understanding.

These previous studies highlight the individual differences among end-users in perceiving AV functionality visualizations and that these could influence their perceptions of safety, trust, acceptance, predictability, and cognitive load [22]. As these perceptions are critical for AVs' public adoption, there is a need to align end-user experiences with AV functionality visualization design.

2.2 Personalization and Computational Methods of In-Vehicle UI Design

Research on automotive UIs over the past ten years [3] and related studies [25] highlight the benefits of personalized in-vehicle UIs that enable tailoring UIs to end-users' perceived safety, trust, and acceptance. For instance, Normark [65] allowed participants to manually personalize icons' size, location, and color on the dashboard, center stack, and HUD. Normark found that participants perceived their custom designs as safer and more usable than a standard design. However, manual personalization by end-users may be impractical. It requires dedicated design settings and is prone to human error if they (unintentionally) create inadequate designs, such as overlapping components or low-contrast colors, endangering driving safety.

Computational methods are another approach for personalization of in-vehicle UIs. These can help designers more effectively choose a design based on subjective designer and/or end-user ratings (e.g., perceived safety, trust, or acceptance). For instance, Zhong et al. [91] derived weights from designers' usability ratings for three HUD design schemes (classic, minimalism, sport). Their computational method used these weights to select the design that best balanced the ratings. However, this method did not incorporate end-users ratings, casting doubt on whether the chosen design fully met their needs and preferences. In contrast, Yunuo et al. [88] allowed end-users to rate HUD design elements such as warn icon style and transparency, each at three levels. These ratings were weighted to select the most end-user-preferred HUD design from 18 predefined options. Although this method identified a design with the highest usability rating among the samples, it potentially overlooks better combinations not included in the 18 samples. Also, it does not refine the design based on iterative end-user feedback. Furthermore, these approaches Yunuo et al. [88], Zhong et al. [91], focused on usability, may not adequately explore the high-dimensional design space of continuous parameter values or address multiple objectives that enhance user experience in AV functionality visualizations.

In contrast, computational optimization methods allow for iterative refinement of designs, aligning more closely with end-users' needs and preferences. Furthermore, they offer a systematic approach to personalization, potentially addressing individual preferences more effectively than manual methods (e.g., see [9]). Despite its potential, research on optimizing in-vehicle UIs through these methods is sparse. Therefore, we use HITL optimization, which integrates designers' expertise and end-users' preferences into an iterative design optimization process. We also consider multiple design objectives, including perceived safety, trust, user acceptance, predictability, and cognitive load, exploring a wider design space with continuous parameter values.

2.3 Human-in-the-Loop Multi-Objective Bayesian Optimization

HITL optimization integrates humans in its iterative parameter optimization cycles when design objectives require humans' subjective ratings (e.g., [11, 50, 79, 90]) or through performance measurements (e.g., [26, 45, 46]).

In Human-Computer Interaction (HCI), BO has been employed in HITL optimization to tackle various design problems [12, 26, 45, 50]. BO is a machine learning method for optimizing unknown and/or difficult-to-evaluate functions [9], such as black-box user models. While there are other black-box optimization methods (e.g., evolutionary and genetic algorithms, see [2]), BO stands out due to its

consistent performance [7] and customizability [56]. BO iteratively evaluates and updates parameters to achieve the best results for a given objective. It balances *exploration*, which involves probing underexplored regions of the design space to discover potentially better designs, and *exploitation*, focusing on areas already identified as promising based on prior knowledge. This balance enables BO to find optimal designs with relatively few iterations, making it one of the most efficient optimization approaches [8]. Therefore, BO is well suited to the problem of AV functionality visualization design, where the relationship between design parameters and user experience is hard to model.

Previous works often use a *cold-start* BO approach (e.g., [9]), where the optimization process relies on initial random sampling to gather data before refining designs. This method, while effective, can be slow due to the lack of prior knowledge [58, 69]. In contrast, Liao et al. [58] have explored a *warm-start* approach, which uses pre-existing data to bypass the sampling phase, leading to faster convergence on an optimal design.

AV functionality visualization design must consider multiple objectives like safety, trust, acceptance, predictability, and cognitive load according to the Automation Acceptance Model [33]. MOBO might be the answer to address these, as it can maximize or minimize multiple objectives simultaneously. The result is not a single optimal design but a range of solutions known as the Pareto front. This front consists of all designs in a multi-objective optimization problem that are not outperformed by any other design. Each point on the Pareto front is *Pareto-optimal*, meaning no other design is better in all objectives simultaneously [61]. It illustrates the best trade-offs between conflicting objectives, where improving one (e.g., usability) may result in a decrease in another (e.g., perceived safety).

MOBO was used, for example, to support the design of touchscreen keyboards that balance speed, user familiarity, and enhanced spell-checking [27], for designing multi-finger inputs for text entry in mid-air [77], for creating haptic interfaces [37], and for an interactive personalization of explanations of image classifier results [10]. These works demonstrated MOBO's effectiveness in HCI design tasks, especially when studies with human participants are costly. Therefore, we argue that MOBO is an appropriate method for optimizing AV functionality visualization design. However, MOBO's effectiveness in this domain and the associated modeling of complex end-user states, such as perceived safety, trust, and acceptance, remains unclear.

Besides, Chan et al. [9] and Liao et al. [57] revealed that designers felt less agency and ownership over MOBO-driven designs, even if they were of higher quality. To foster collaboration between BO-supported design approaches and designers, Koyama and Goto [49] introduced BO as a design assistant. This allowed designers to tap into their expertise and preferences while BO offered design suggestions.

In contrast, involving end-users in AV functionality visualization design is critical, as they offer key insights into their preferences that designers cannot have. While previous research primarily explored the impact of designers in HITL MOBO design processes (and vice versa), end-user integration is left underexplored. Yet, their contributions can potentially enhance optimization efficiency and design quality compared to designers. Therefore, this work focuses on engaging end-users in the design of visualizations to address existing HITL MOBO limitations. Besides, we extend the understanding of qualities of HITL optimization established in works like [9, 57] by comparing end-user-led and optimizer-driven processes.

3 OptiCarVis: Optimizing Automated Vehicle Functionality Visualizations

Visualizations on WSDs use icons [59], highlighting [14, 43], and other elements (e.g., see [52]) to communicate AV functionalities, which could be crucial for user acceptance [14, 43]. These visualizations aim to increase end-users' perceived safety and trust in AVs, their understanding of AV actions in various driving situations (e.g., unexpected stopping) [43], and reduce the cognitive load when processing this (potentially overwhelming) information.

We aim to optimize the visualization design computationally using a HITL MOBO approach. In this work, we solely focus on AR WSD visualizations as these allow situated visualizations, which are beneficial for understanding and reducing cognitive load [13]. By involving end-users, the HITL MOBO process can iteratively refine designs to cater to individual preferences within a reasonable timeframe. Our visualizations (see Figure 2 and Figure 3) are grounded in prior research investigating trust, cognitive load, and perceived safety in AVs (see Section 2.1). We primarily built upon the work of Colley et al. [20], who visualized the functional levels of AVs' internal operations (see [24]): *Situation Detection*, *Situation Prediction*, and *Trajectory Planning*.

Within this framework, Situation Detection is encoded via semantic segmentation of detected objects (i.e., detected vehicles are colorized in blue, pedestrians in red, and traffic signs in yellow, see Figure 2 a). Situation Prediction is encoded via showing the pedestrian intentions symbolized as icons above pedestrians' heads [13] (i.e., the color coding indicates the prediction whether pedestrians are to cross the street (dark blue), whether they will remain on the sidewalk (cyan), or whether the prediction is uncertain (yellow); see Figure 2 b), and deduced trajectories of other vehicles are shown as a line. Here, the color changes from blue to red the further the prediction lies in the future to visualize the increasing uncertainty. The Trajectory Planning [20] (i.e., the planned trajectory of the own vehicle) is also visualized via this line (see Figure 2 c and d). Furthermore, we incorporated elements from Müller et al. [64], focusing on CAD visualizations. CAD supports the AV in all three framework stages and includes additional information as blue spheres above the road, indicating the AV's active link to external sensors in a given area (see Figure 2 e). We also include an outline representing occluded cars through buildings (see Figure 2 f). This visualization supports the end-user in understanding that the AV knows about other vehicles even if they are hidden, for example, by buildings, which may be unintuitive [64]. Finally, a vehicle status HUD displays the current time, AV speed, and destination as basic information (see Figure 2 g).

We define the optimization of AV functionality visualization design as the task of finding the design parameter combination $x* = \{p_1, \ldots, p_n\}$ such that:

$$x* = \arg\max_{x \in X} f(x) \tag{1}$$

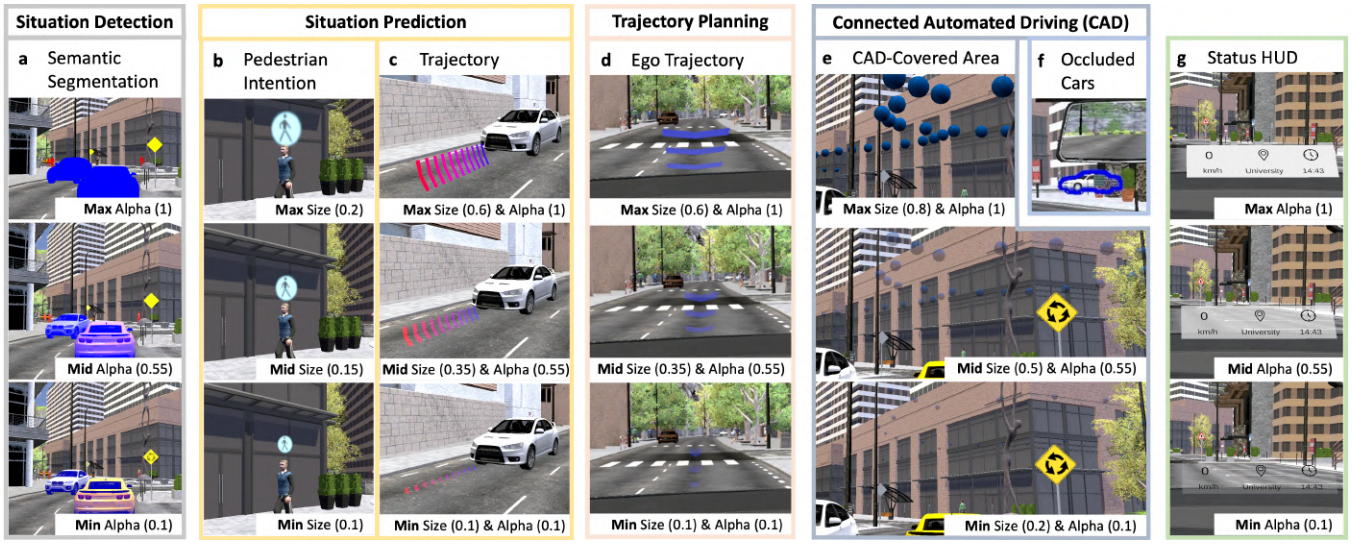


Figure 2: Overview of the employed visualizations of an SAE Level 4 [72] AV's functional levels of internal operation [20], CAD [64], and status on an AR WSD, showing the possible variations in transparency (alpha) and size values (see brackets). *Min* and *Max* represent the designs at the lower and upper bounds of the continuous parameter ranges, while *Mid* represents the midpoints.

where $X \subseteq \mathbb{R}^n$ is the design space defined by n design parameters $p \in \mathbb{R}$. The objective function $f : X \rightarrow Y \subseteq \mathbb{R}^m$ maps each design x to m subjective metrics (e.g., trust and perceived safety). An objective function value $y \in Y$ is a subjective metric rating (e.g., via Likert scale) the end-user returns to the MOBO in the HITL process after viewing a visualization design x (see Figure 1). As the relationship between x and y is unclear, we define this as a black-box function $y = f(x)$ [2].

3.1 Design Parameters

The design parameters for the visualizations were derived from the respective publications [13, 20–22, 51, 64, 71]. As visualization elements might be unwanted, we defined the visualization visibility v per element as $v \in [0, 1]$. Specifically, we set a threshold to map v as a Boolean. An element is invisible for $v < 0.5$ and visible for $v \geq 0.5$. We employed this mapping as BO is typically more efficient with continuous parameters [75]. This Boolean value was determined for the semantic segmentation, (ego) trajectory, pedestrian intention, highlighting occluded cars, the CAD-covered area, and the vehicle status HUD. Besides, element size may denote importance and determine far-distance visibility. Therefore, we added the size s as a parameter of the pedestrian intention icon, (ego) trajectory, and CAD-covered area sphere. We defined s within a range where the bounds indicate the smallest and the largest appropriate size. These bounds were different for each element so that they can neither be too small and thus invisible nor too large and overlap with other elements. Due to AR WSD elements potentially overlaying the driving environment, like other vehicles, their semi-transparency might increase the visibility of the environment. We assigned an alpha level α to the semantic segmentation, (ego) trajectory, CAD-covered area sphere, and vehicle status HUD. The range was $\alpha \in [0.1, 1]$ as elements become nearly invisible for $\alpha < 0.1$. The "occluded cars" visualization does not incorporate an alpha value because it does not block relevant visual information ("occluded cars" is a simple outline). Besides, we did not assign an α level to the pedestrian intention icon to avoid confusion between color coding because semi-transparent dark blue (likely to cross) looks similar to cyan (remaining on the sidewalk).

We avoided RGB coloring as parameters as the color was already chosen not to convey unintended meaning (e.g., orange being a warning signal; e.g., see [20, 64]). We also refrained from altering the position of the CAD-covered area spheres (e.g., via height) and the vehicle status HUD (e.g., x and y position on the windshield) as the proposed constant positions are the most useful. During optimization, these constant positions prevent visualization elements' misalignment (e.g., overlapping) and ensure visibility for multiple passengers' viewpoints. Similarly, we disregard the visualization elements' rotations as these are already determined by the objects' orientations in the environment, such as vehicles and pedestrians. All design parameters (p_1 to p_{16}) are summarized in Table 1.

3.2 Objective Functions

An objective function f maps a visualization design x to a subjective metric the optimizer seeks to maximize or minimize with the design. According to our optimization goal, we consider five subjective metrics - *safety*, *trust*, *predictability*, *acceptance*, and *aesthetics* - to be maximized. *Cognitive load* was our sole subjective metric to be minimized.

Based on previous work [13, 20], we employed the following questionnaires to retrieve these metrics after every optimization iteration in the HITL process: We assessed **cognitive load** via the mental workload subscale of the raw NASA-TLX [35] on a 20-point scale ("How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex?"; 1=*Very*

Table 1: The 16 design parameters for the visualization design, with the ranges in Unity values. All design parameters are modeled continuously, with values mapped to Boolean if necessary ("Bool"). Example visualizations of parameter values are shown in Figure 2.

Design Parameter	Description	Reference	Range
p_1: Semantic Segmentation, v	Whether the semantic segmentation result should be visualized.	[14]	[0, 1]; Bool
p_2: Semantic Segmentation Alpha, α	Alpha value of the semantic segmentation.	[14]	[0.1, 1]
p_3: Pedestrian Intention, v	Whether the predicted pedestrian intention should be visualized.	[13]	[0, 1]; Bool
p_4: Pedestrian Intention Size, s	Alpha value of the pedestrian intention symbol.	[13]	[0.1, 0.2]
p_5: Trajectory, v	Whether the predicted trajectory of others should be visualized.	[21, 51]	[0, 1]; Bool
p_6: Trajectory Alpha, α	Alpha value of the trajectory.	[51]	[0.1, 1]
p_7: Trajectory Size, s	Size of the trajectory.	[51]	[0.1, 0.6]
p_8: Ego Trajectory, v	Whether the own planned trajectory should be visualized.	[20, 21]	[0, 1]; Bool
p_9: Ego Trajectory Alpha, α	Alpha value of the own planned trajectory.	[20, 21]	[0.1, 1]
p_{10}: Ego Trajectory Size, s	Size of the own planned trajectory.	[51]	[0.1, 0.6]
p_{11}: CAD-Covered Area, v	Whether the area covered through V2x should be visualized.	[64]	[0, 1]; Bool
p_{12}: CAD-Covered Area Alpha, α	Alpha value of the symbols for the CAD-covered area.	[64]	[0.1, 1]
p_{13}: CAD-Covered Area Size, s	Size of the symbols for the CAD-covered area.	[64]	[0.2, 0.8]
p_{14}: Occluded Cars, v	Whether occluded (e.g., by buildings) cars should be visualized.	[64]	[0, 1]; Bool
p_{15}: Vehicle Status HUD, v	Whether the vehicle status in the HUD should be visualized.	[22]	[0, 1]; Bool
p_{16}: Vehicle Status HUD Alpha, α	Alpha value of the vehicle status.	[71]	[0.1, 1]

Low to 20=*Very High*; lower is better). Regarding predictability and trust, we used the subscales *Predictability/Understandability* (*Predictability*) and *Trust* of the *Trust in Automation* questionnaire by Körber [48]. **Predictability** is determined via agreement on four statements ("The system state was always clear to me.", "I was able to understand why things happened."; two inverse: "The system reacts unpredictably.", "It's difficult to identify what the system will do next.") using 5-point Likert scales (1=*Strongly disagree* to 5=*Strongly agree*). **Trust** is measured via agreement on the same 5-point Likert scale on two statements ("I trust the system." and "I can rely on the system."; both times, higher is better). Participants rated their perceived **safety** using four 7-point semantic differentials from -3 (anxious/agitated/unsafe/timid) to +3 (relaxed/calm/safe/confident; higher is better) [29]. Finally, we added three single items. Two were defined with the van der Laan acceptance scale [80] in mind ("I find the visualizations of the automated vehicle **useful**", "I find the visualizations of the automated vehicle **satisfying**"). These were combined into a single "acceptance" objective. We also adapted the question regarding **aesthetics** from Colley et al. [19] ("I found the visualizations visually appealing"; on a 7-point Likert scale).

Normalization is required before submitting these to the optimizer because the subjective metrics values have ranges based on 20-, 5-, or 7-point Likert scales. We transformed these six metrics into the $[-1, 1]$ range. After the transformation, the *cognitive load* objective is a function to be maximized (a higher value means less load).

3.3 Hyperparameter Setup for Bayesian Optimization

For our MOBO implementation, we used the PyTorch-based library `BoTorch` [4] in version 0.9.2. As we have a multi-objective setup, we employed the multi-output Gaussian Process and applied `qEHVI` as the acquisition function. This function represents the expected hypervolume increase, where we set q = 1 (in line with [9]) to ensure that after each iteration, a batch of size one is selected to be given to the end-user for evaluation. Other hyperparameter settings were 5 sampling iterations followed by 10 optimization iterations. During the optimization of the acquisition function, 2024 restart candidates for the acquisition function optimization, and 512 Monte Carlo samples were used to approximate the acquisition function. These settings are based on Chan et al. [9].

In internal tests, we found that the convergence to an optimal rating of the objectives was reached rather quickly. Therefore, we added a stopping criterion checked after every measurement: Was the perfect rating for **every** subjective metric (i.e., the **highest** rating for trust, predictability, safety, aesthetics, usefulness, satisfaction, and the **lowest** rating for cognitive load; see Section 3.2) given for the **last** round? Participants could otherwise not opt out of the optimization steps.

4 Experiment

We aim to empirically validate the effectiveness of the HITL MOBO approach for designing AV functionality visualizations, comparing it with traditional manual designs. Building on Chan et al. [9], we investigate the mutual influence between the HITL optimization and its end-users during simulated automated driving. Additionally, we expect varied outcomes when initializing the MOBO (i.e., Warm-Start) with data from automotive UI experts or end-users. Guided by these goals, we conducted a user study with the following research questions (RQs):

- **RQ1** How does the HITL MOBO of AV functionality visualizations impact end-users' rating of safety, trust, predictability, acceptance, aesthetics, and cognitive load?
- **RQ2** Which condition produces the design leading to the highest rating of safety, trust, predictability, acceptance, aesthetics,

and the lowest cognitive load: Cold-Start with random initial parameters, Warm-Start initialized by expert designs, or Warm-Start initialized by end-user designs?

RQ3 How does the participation in a HITL optimization process affect end-users, and what are their areas of interest during design?

The experimental procedure followed the guidelines of our university's ethics committee and adhered to regulations regarding handling sensitive and private data, anonymization, compensation, and risk aversion. Compliant with our university's local regulations, no additional formal ethics approval was required.

4.1 Apparatus

The apparatus comprises three main components: (1) a Unity application for the driving environment, (2) a custom parameter design tool, and (3) a Bayesian optimizer. We also used a server to persistently store the local questionnaire responses persistently, the design parameter logs from optimization rounds, participants' ratings, and optimization durations.

4.1.1 Automated Vehicle and Driving Environment. We designed a standalone application for Windows and macOS using Unity 2022.3.7. This application simulates in-vehicle visualizations in a driving environment using a 3D model of the Tesla Model X, modified to feature a virtual AR WSD and a vehicle status HUD. As participants should not overtake control but expect errors in automation, we consider a SAE Level 4 AV [72]. The driving environment, Unity Windridge City, aligns with previous research [13, 14, 20]. We used the Urban Traffic System asset to simulate pedestrian and vehicle behaviors. Each MOBO iteration employs a fixed 33-second route. This brief duration is chosen to reduce user fatigue during the HITL process. We also developed a longer 3-minute route to provide users with a broader range of traffic situations. Both routes (see Figure 3) incorporate frequent pedestrian and vehicle interactions at roundabouts and zebra crossings, creating diverse visualization scenarios.

4.1.2 Custom Parameter Design Tool. In Unity, we developed a tool (see Figure 4) that enables adjustment of the 16 parameters (see Table 1) for custom visualization designs. Users can toggle the element visibility v using a checkbox and adjust the element transparency α and size s using sliders within the predefined parameter value ranges. A side-by-side preview panel continuously displays the AV driving environment with the current settings. As users modify parameters, this environment loops. Once users finalize their settings, they confirm via a button, saving the parameter value configuration as initial data for the Bayesian optimizer.

4.1.3 The Bayesian Optimizer. During the HITL optimization, the Bayesian optimizer interacts with the Unity application. It iteratively receives user ratings for the current visualization design (the optimization objectives, see Section 3.2) and returns the next potentially optimal parameters in CSV format. To guarantee prompt computation, the optimizer runs locally on participants' computers. The configuration of this optimizer for in-vehicle visualization design is detailed in Section 3.3.

4.2 Design and Optimization Conditions

To answer RQ1 - RQ3 (see Section 4), we employ the following conditions, building upon prior research [9, 57]:

- **C1 No Visualization (No Vis.):** In this condition, no visualization of AV functionalities is shown.
- **C2 Custom design by experts:** End-users evaluate a *standard* design created by automotive UI experts (N=8, see Section 4.3) using our parameter design tool (see Section 4.1.2 and Table 2). This visualization uses the mean parameter values from all expert designs. Unlike C5, this condition may result in suboptimal designs, as a "one-size-fits-all" *standard* expert design alone may not optimally adhere to the individual subjective ratings of end-users.
- **C3 Custom design by end-users:** Similar to C2, but instead of evaluating a *standard* design, end-users manually personalize visualizations using our parameter design tool and evaluate them after the AV ride. Unlike C6, this condition may result in suboptimal designs, as end-users may not fully understand their preferences and ineffectively translate them in a direct parameter design process.
- **C4 Cold-Start HITL MOBO:** We use a Cold-Start HITL MOBO initialized with random parameters generated by the optimizer. The MOBO starts in the *sampling* phase. End-users then interactively rate potential designs, fed back into the optimizer in a HITL process.
- **C5 Expert-Informed Warm-Start HITL MOBO:** As in C2, we enable automotive UI experts (N=8, see Section 4.3) to explore designs using our parameter design tool. End-users then rate this *standard* expert design, which creates design objective values for the given parameters. In the Warm-Start approach, these values initialize the HITL MOBO. Thus, the *sampling* phase is replaced, and it directly starts with the *optimization* phase, in which the end-user is iteratively involved. This condition leverages experts' domain knowledge to focus the design space on areas likely aligned with automotive UI best practices. It potentially optimizes designs by further personalizing a given design foundation to fit end-users' subjective ratings.
- **C6 User-Informed Warm-Start HITL MOBO:** End-users first explore personalized visualization designs using our parameter design tool. After that, they rate their design to generate the objective values to initialize the HITL MOBO *optimization* phase. This Warm-Start HITL MOBO process then fine-tunes the user-informed parameter values, combining end-users' preferences with optimization to uncover optimized designs when they cannot fully express their needs in a custom design. Like in C5, this potentially accelerates MOBO's discovery of optimal designs as end-user preferences narrow the design space [9, 57]. Besides, such Warm-Start HITL MOBO augmentation could combat users' feelings of low agency in HITL approaches [9].

Only C4-C6 represent personalization methods that include HITL MOBO. C1-C3 represent the current state of the art by not showing any information or letting experts or end-users define the visualization themselves. We exclusively employ MOBO as our black-box optimization method due to its consistently robust performance

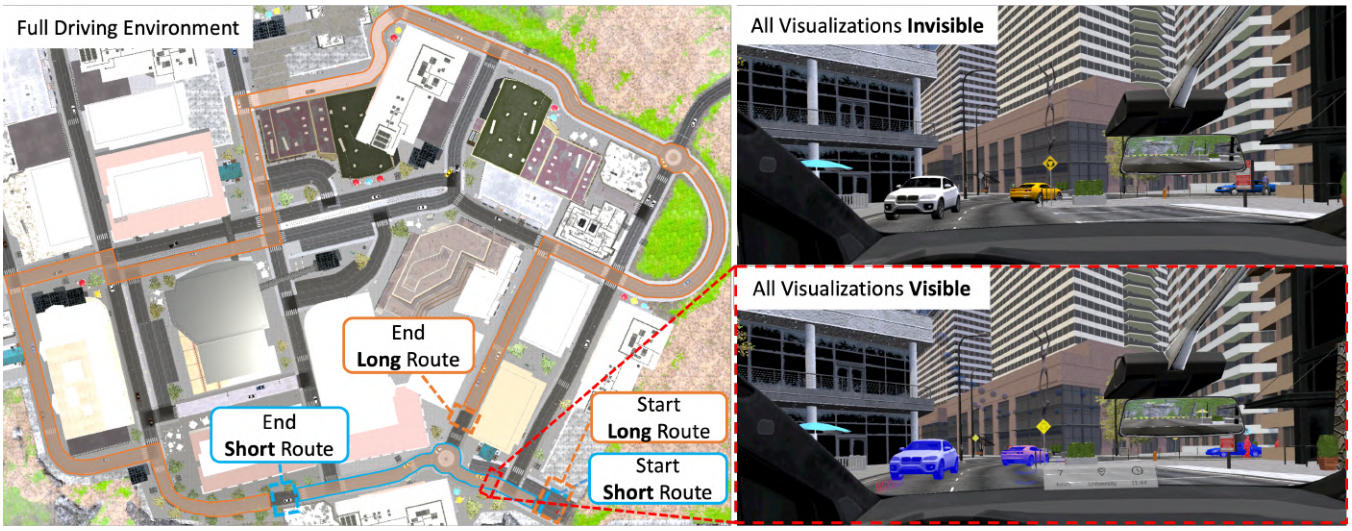


Figure 3: AV study driving route used in the HITL MOBO iterations (blue) and long route used in the final assessments (orange). Besides, examples of the driver's perspective with all visualizations visible using *mid* transparency and size values (red).

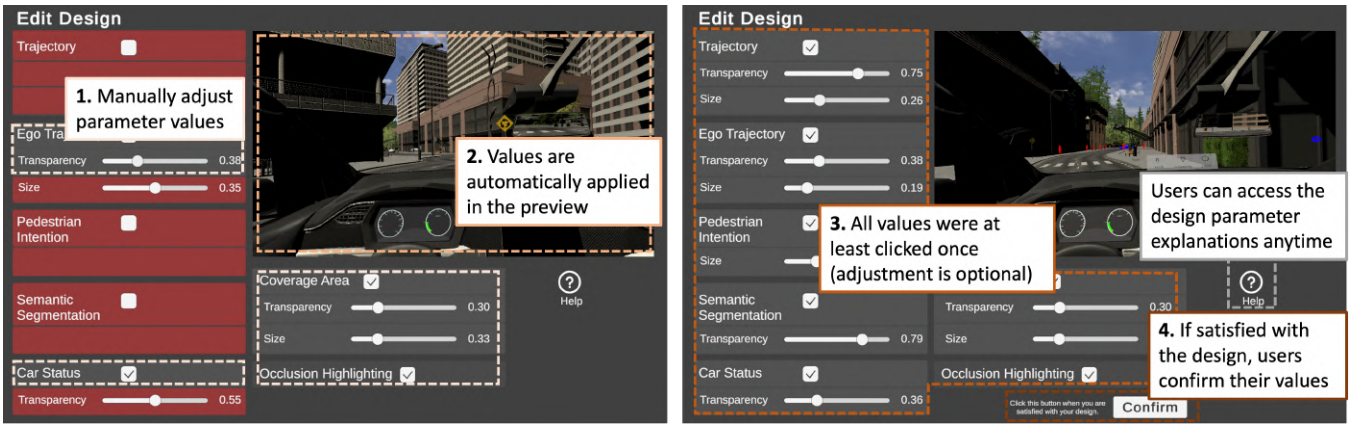


Figure 4: The custom parameter design tool allows for adjusting 16 parameters (see Table 1). (1) Users modify values using checkboxes and sliders, with untouched settings highlighted in red. (2) The adjusted values are displayed in a preview that loops the AV driving environment. (3) After interacting with all settings once (their adjustment is optional), (4) the "confirm" button activates. The parameter explanation view (see Figure 5) is accessible via the "help" button.

across applications [7], unparalleled customizability tailored to the unique requirements of HCI design [56], and superior efficiency in converging on optimal designs [8]. Therefore, we argue that including competitor optimization methods would unlikely add value to our experiment. We define the following hypotheses:

- **H1** HITL MOBO for AV functionality visualizations (C4-C6) will increase end-users ratings of safety, trust, predictability, acceptance, and aesthetics and reduce cognitive load compared to non-MOBO conditions (C1-C3).
- **H2** Among the HITL MOBO conditions, the C6-User-Informed Warm-Start will result in higher ratings for safety, trust, predictability, acceptance, aesthetics, and lower cognitive load, outperforming both the C4-Cold-Start and the C5-Expert-Informed Warm-Start.

4.3 Expert Study to Inform the Standard Visualization Design

For conditions C2 and C5, addressing RQ2, we aimed to create a *standard* visualization design using expert insights on typical automotive UI design practices. We recruited N=8 automotive UI experts (2 female, 6 male, 0 non-binary) who specialized in in-vehicle UI usability and trust in automation. These experts, with backgrounds in psychology (1), computer science/HCI (6), and engineering (1), represented four institutions from Europe, the USA, and Canada. They hold positions as research associates and Ph.D. students or are currently or were engineers at two large European OEMs. Participants were, on average, M=27.88 (SD=2.36) years old. All have

published multiple papers on automotive design. Publishing in automotive design-oriented venues constitutes *Experience* and *Peer Identification*, which, in the sense of Shanteau et al. [76], constitute experts. The participants' expertise in designing **and** evaluating automotive UIs regarding subjective end-user ratings makes them a perfect fit for this study. The experts were tasked to design AR WSD visualizations for AVs that enhanced end-users' perceived safety, trust, predictability, and acceptance of AVs while reducing cognitive load.

Each session started with a brief instruction on the available visualizations (see Figure 5) and parameters (see Table 1), informed consent, and a demographic questionnaire. Using the custom design tool in Unity (see Section 4.1.2 and Figure 4), experts freely adjusted the 16 design parameters. By default, the visibility checkboxes were set to false, while the sliders were set to the mid-value. The preview panel continuously visualized the current configuration in a looped scene, enabling experts to refine their designs iteratively. Once satisfied, they confirmed their parameter configuration and answered open-ended questions about design rationales. The resulting parameter values are shown in Table 2. For C2 and the initialization of the HITL MOBO in C5, we used a *standard* visualization derived by averaging each parameter from the expert designs. The averaging of expert opinions avoids bias towards any viewpoint and balances contrasting parameters (e.g., a trajectory alpha of one compared to 0.31).

To verify the averaged design, three authors individually reviewed and then collaboratively discussed it, supplemented by qualitative comparisons to the previous works from which we derived our visualizations [13, 20, 22, 51, 64, 71]. Most design parameters exhibited relatively limited variability ($SD \leq 0.18$; see Table 2). Specifically, Semantic Segmentation (p_1, p_2), Pedestrian Intention (p_3, p_4), Trajectory (p_5, p_7), Ego Trajectory Alpha (p_9) and Size (p_{10}), CAD-Covered Area Size (p_{13}), and Vehicle Status HUD (p_{15}) had $SD \leq 0.18$, indicating a fair degree of alignment among experts. In contrast, six parameters had more pronounced variability ($SD > 0.18$), suggesting diverse views. These included Trajectory Alpha (p_6), Ego Trajectory (p_8), CAD-Covered Area (p_{11}, p_{12}), Occluded Cars (p_{14}), and Vehicle Status HUD Alpha (p_{16}). Post-study interviews indicated that personal preferences influenced alpha parameter choices, making it challenging for experts to settle on a general design suited for all users. However, the comparatively lower variability across most parameters suggests the averaged design parameters are reasonable. After reviewing the final averaged design, all experts noted it fell within an acceptable range of values. This combination of quantitative averaging and qualitative insights aligns with established industry practices for consensus building in empirical A/B studies [62].

4.4 Participants

We computed the desired sample size for the main experiment via an a-priori power analysis using G*Power [31]. To achieve a power of 0.95, with an alpha of 0.05, 111 participants should allow for detection of a medium effect (*Effect Size f*=0.25) in repeated measures ANOVA with the conditions C1-C6 as a between-subject factor.

Thus, we recruited 117 participants (Mean age = 39.2, SD = 12.3, range: [19, 72]; Gender: 29.9% women, 68.4% men, 1.71% non-binary; Education: College, 72.65%; High School, 20.51%; Vocational training, 6.84%) via prolific.co. C1 had 21 participants, C2 19, C3 19, C4 18, C5 22, and C6 18.

To prevent confounding effects of traffic handedness (right-hand vs. left-hand traffic) or culture, the participant pool was limited to US residents [70] as the Unity simulation employed right-handedness. Regarding their employment status, 81 are employees, eight are students at a college, one is at a school, 15 are self-employed, 9 are job-seeking, and three indicated *other*. All participants hold a valid driver's license for, on average, M=18.62 (SD=13.18) years. We found no significant differences between the *visualization condition* for license, gender, or age. All volunteered under informed consent and agreed to the recording and anonymized publication of results. Participants were compensated with £7.

4.5 Procedure

We conducted the study online to engage diverse end-users with non-technical backgrounds, a typical challenge in lab settings. Moreover, safely simulating AVs in a computer-screen-based Unity application is a widely adopted method for evaluating novel in-vehicle UIs (e.g., see [13, 20, 44]) as such an AV technology is not yet available. Using a between-subjects, the study employed the distinct conditions **C1-C6**. The baseline condition (C1-No Vis.), which displayed no visualizations during the AV ride, was added to validate the effectiveness of designs from C2-C6.

Participants were distributed across the **six** conditions (see Figure 6). Upon downloading the Unity driving simulation and the Bayesian optimizer (see Section 4.1), sessions began with a short introduction, informed consent, and a demographic questionnaire. The AV introduction was adapted from Colley et al. [20].

For conditions C2-C6, participants were informed that the AV would display detected objects, their predicted actions, and the AV's planned maneuvers on its WSD (see Figure 5). Additionally, they received a brief overview of the visualizations with examples for semantic segmentation, pedestrian intention, trajectory, ego trajectory, CAD-covered area, occluded cars, and status HUD (see Figure 2 a-g). Despite the visualizations performing perfectly (i.e., always highlighting all relevant objects), participants were intentionally informed that the AV "attempts to assess the situation," implying the possibility of errors during the ride. This introduced a sense of potential risk to establish subjective rating levels (e.g., trust, see [55]).

According to our optimizer setup (see Section 3.3), the 33-second driving route (see Figure 3) was repeated 15 times (5 sampling and 10 optimization iterations) in C4 and 10 times in C5 and C6 (only optimization due to the available data in C5 and C6 which were used for the sampling phase). For the final user rating, we employed the 3-minute route to ensure the final designs were experienced in new situations. In non-MOBO conditions C2 and C3, participants only experienced the 3-minute route. In C4-C6, they undertook multiple trips in the AV on the same route. The total task time was 3 minutes for C1 and C2, up to 12 minutes for C3 (including up to 8 minutes of custom designing), 11.25 minutes for C4, and 8.5 minutes for C5 and C6.

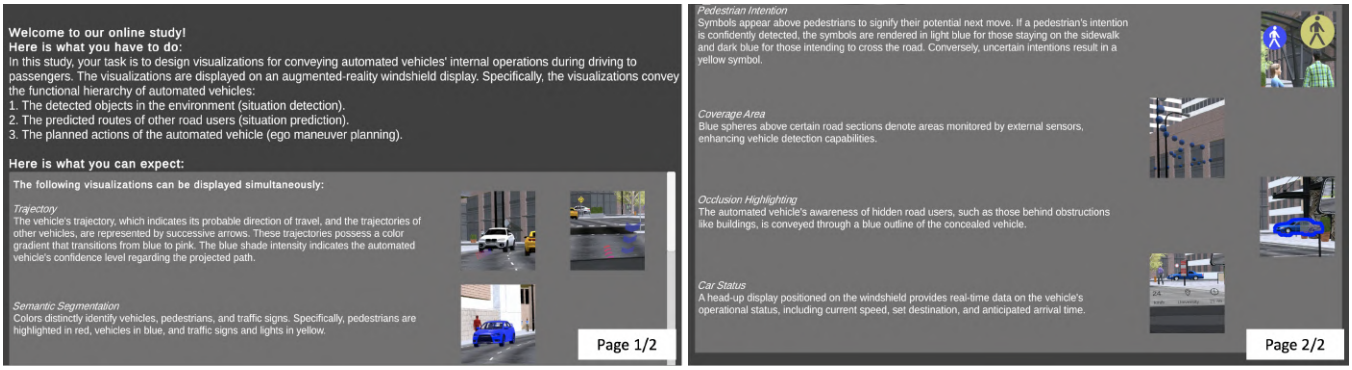


Figure 5: Excerpt of the information given to study participants at the start. Participants were also questioned about the visualizations to ensure understanding.

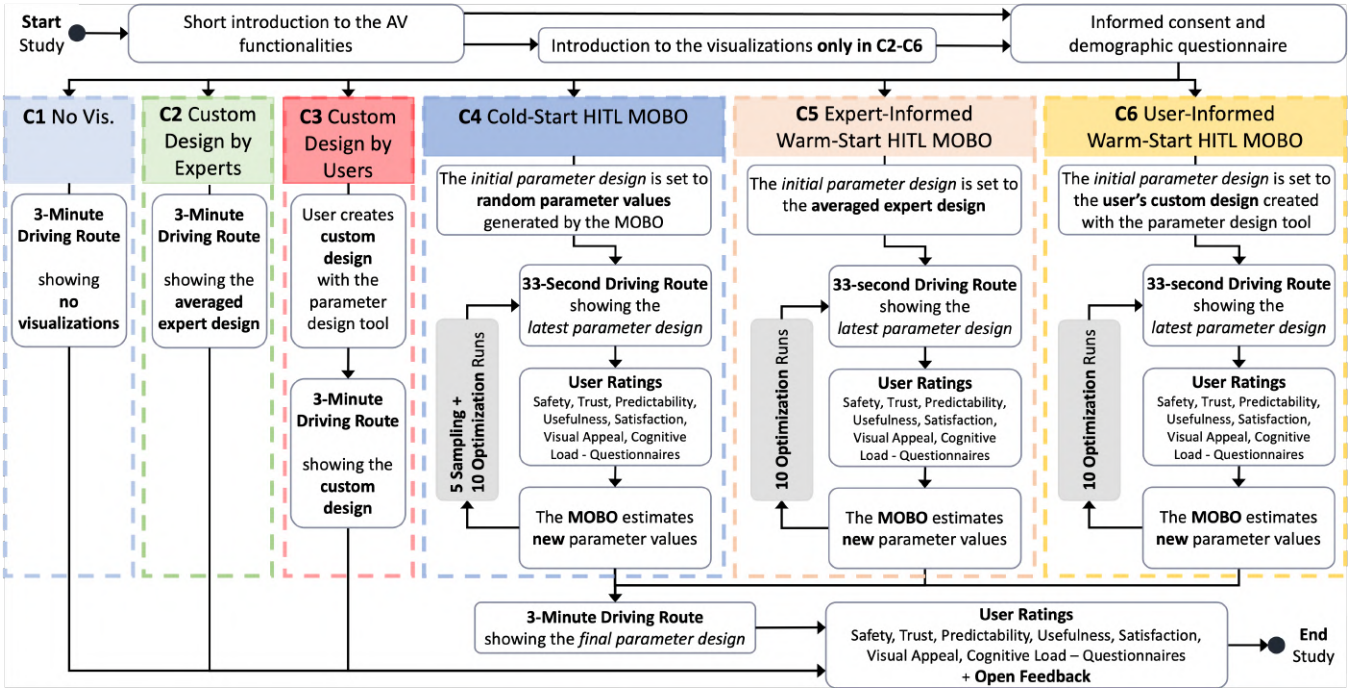


Figure 6: Study procedure of the six conditions C1-C6.

We did not inform participants they were part of a HITL optimization process or detail how the optimizer applied their feedback to the design. In real-world scenarios, especially with in-vehicle interactions, we argue that users do not have deep knowledge about the system's operation but can still evaluate the quality of an experience. In total, the study lasted up to 50 minutes, and we integrated attention and comprehension checks following Prolific's guidelines. Participants could not intervene in the driving task as the visualizations are primarily intended to inform the user about AV functionalities.

4.5.1 Subjective Ratings. During the HITL optimization, participants rated the visualization designs via the subjective metrics defined in Section 3.2 **after** each ride, with the possibility for textual feedback. For the C1-*No Vis.* condition, we did not assess Acceptance (i.e., Usefulness and Satisfaction [80]) and Aesthetics as these would not make sense without any visualization. After the session, in the final user rating, we measured the subjective metrics defined in Section 3.2 and the design experience using adapted questions from Chan et al. [9]. On 7-point Likert scales (1=*Strongly disagree* to 7=*Strongly agree*), we queried about *User Expectation Conformity*: "The final design matches my imagination.", *Satisfaction*: "I'm pleased with the final design.", *Confidence*: "I believe the design is optimal for me.", *Agency*: "I felt in control of the design process." and *Ownership*: "I feel the final design is mine." Regarding *Interactivity*, participants also provided feedback on desired design control levels ("... Consider aspects where you desired more or less control

over the design."). Participants could further elaborate with textual comments if they disagreed with the statements for expectation or satisfaction.

4.5.2 Objective Measures. We recorded the Bayesian optimizer's performance metrics and the time taken in the Unity application for questionnaire responses (C2-C6) and custom design (C2 and C6). For RQ3, we embedded the webcam-based eye-tracker UnitEye[2] in the study application to track areas of interest (AOIs): pedestrian, vehicle, traffic sign, pedestrian intention icon, occluded car, CAD-covered area sphere, and vehicle status HUD. Participants calibrated the eye-tracker before the study so that we could monitor their focus and attention during the design and the AV ride on the 33-second and 3-minute driving routes. The eye-tracking data was not used as an objective function of the Bayesian optimizer.

5 Results

5.1 Quantitative Results

5.1.1 Data Analysis. Before every statistical test, we checked the required assumptions (e.g., normality distribution). R in version 4.4.2 and RStudio in version 2024.09.0 were employed. All packages were up-to-date in December 2024. We used the ARTool package by Wobbrock et al. [86] for non-parametric data as the typical ANOVA is inappropriate with non-normally distributed data and Holm correction for post-hoc tests. The procedure is abbreviated with ART. For the comparisons report in Section 5.1.3, we used **all Pareto front** values per MOBO condition per user. Figure 7 to Figure 9b show only significant differences via bars using Dunn's test for post-hoc comparisons with Holm correction. The progression of the dependent variables **during** the MOBO iterations are shown in Figure 12, Figure 13, and Figure 14.

The error bars represent bootstrap confidence intervals (i.e., mean_cl_boot). We refrain from reporting the Pareto front graphically due to (1) the high number of parameter value combinations—even with only five discrete levels for our nine continuous design parameters (four s and five α values, see Table 1), there are approximately 5^9 possible combinations—and (2) the resulting challenges in visualization due to the high number of dimensions ($p_1 \dots p_{16}$).

5.1.2 Number of Applied Stopping Criterion. The stopping criterion was met when all six design objectives received perfect scores in two consecutive iterations (see Section 3.3). This occurred for 12 of the 57 participants (21.05%) interacting with a HITL MOBO variant: three in C4-Cold-Start, four in C5-Expert-Informed Warm-Start, and five in C6-User-Informed Warm-Start HITL MOBO. Achieving perfect (i.e., maximum/minimum) scores across multiple objectives is challenging, making the 21.05% proportion notable. Besides, the trend lines in Figure 12 to Figure 14 and the high scores of the other participants suggest that additional iterations would result in more users reaching perfect scores. We interpret this as validation of our stopping criterion and the assumption that convergence occurs quickly.

5.1.3 Design Performance. A Kruskal-Wallis rank sum test found a significant effect of *visualization condition* on *perceived safety* ($\chi^2(5)=39.44$, $p<0.001$, r=0.11; see Figure 7a).

A post-hoc test found that C4-Cold-start HITL MOBO was significantly higher ($M=1.32$, $SD=1.72$) in terms of *perceived safety* compared to C2-Custom design by experts ($M=-0.25$, $SD=1.55$); *adj.* $p<0.001$), compared to C3-Custom design by end-users ($M=-0.70$, $SD=1.97$); *adj.* $p<0.001$).

A post-hoc test also found that C5-Expert-Informed Warm-Start HITL MOBO was significantly higher ($M=0.95$, $SD=1.99$) in terms of *perceived safety* compared to C2-Custom design by experts ($M=-0.25$, $SD=1.55$); *adj.* $p=0.009$) and compared to C3-Custom design by end-users ($M=-0.70$, $SD=1.97$); *adj.* $p<0.001$).

A post-hoc test finally also found that C6-User-Informed Warm-Start HITL MOBO was significantly higher ($M=0.74$, $SD=1.62$) in terms of *perceived safety* compared to C3-Custom design by end-users ($M=-0.70$, $SD=1.97$); *adj.* $p=0.014$).

A Kruskal-Wallis rank sum test found a significant effect of *visualization condition* on cognitive load ($\chi^2(5)=30.17$, $p<0.001$, r=0.08; see Figure 7b). A post-hoc test found that C2-Custom design by experts was significantly higher ($M=9.76$, $SD=4.72$) in terms of *cognitive load* compared to C4-Cold-start HITL MOBO ($M=6.71$, $SD=3.76$); *adj.* $p=0.008$), compared to C6-User-Informed Warm-Start HITL MOBO ($M=6.72$, $SD=4.31$); *adj.* $p=0.008$), and compared to C5-Expert-Informed Warm-Start HITL MOBO ($M=6.12$, $SD=4.05$); *adj.* $p<0.001$). A post-hoc test also found that C3-Custom design by end-users was significantly higher ($M=9.03$, $SD=3.64$) in terms of *cognitive load* compared to C4-Cold-start HITL MOBO ($M=6.71$, $SD=3.76$); *adj.* $p=0.032$), compared to C6-User-Informed Warm-Start HITL MOBO ($M=6.72$, $SD=4.31$); *adj.* $p=0.032$), and compared to C5-Expert-Informed Warm-Start HITL MOBO ($M=6.12$, $SD=4.05$); *adj.* $p=0.003$).

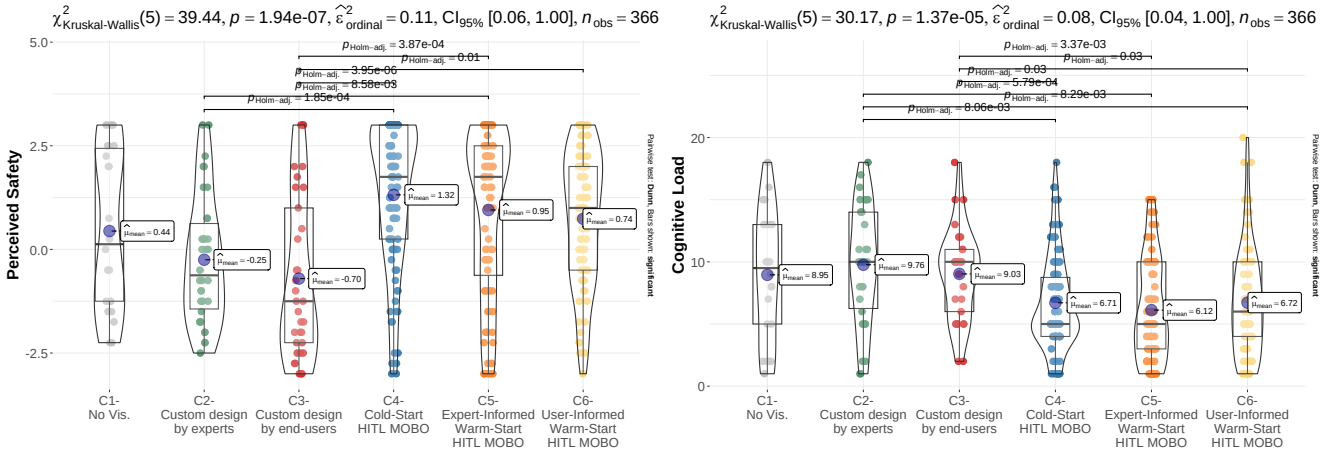
A Kruskal-Wallis rank sum test found a significant effect of *visualization condition* on trust ($\chi^2(5)=24.42$, $p<0.001$, r=0.07; see Figure 8a). A post-hoc test found that C4-Cold-start HITL MOBO was significantly higher ($M=3.94$, $SD=1.07$) in terms of *trust* compared to C2-Custom design by experts ($M=3.04$, $SD=1.17$); *adj.* $p<0.001$). A post-hoc test also found that C5-Expert-Informed Warm-Start HITL MOBO was significantly higher ($M=3.95$, $SD=0.89$) in terms of *trust* compared to C2-Custom design by experts ($M=3.04$, $SD=1.17$); *adj.* $p=0.002$).

A Kruskal-Wallis rank sum test found a significant effect of *visualization condition* on predictability ($\chi^2(5)=32.55$, $p<0.001$, r=0.09; see Figure 8b). A post-hoc test found that C4-Cold-start HITL MOBO was significantly higher ($M=4.03$, $SD=1.00$) in terms of *predictability* compared to C2-Custom design by experts ($M=3.10$, $SD=1.18$); *adj.* $p<0.001$), compared to C1-No Vis. ($M=3.08$, $SD=1.03$); *adj.* $p=0.001$), and compared to C3-Custom design by end-users ($M=3.38$, $SD=1.17$); *adj.* $p=0.021$). A post-hoc test also found that C5-Expert-Informed Warm-Start HITL MOBO was significantly higher ($M=3.84$, $SD=0.89$) in terms of *predictability* compared to C2-Custom design by experts ($M=3.10$, $SD=1.18$); *adj.* $p=0.021$), compared to C1-No Vis. ($M=3.08$, $SD=1.03$); *adj.* $p=0.041$)

A Kruskal-Wallis rank sum test found no significant effects on acceptance ($\chi^2(4)=8.58$, $p=0.073$, r=0.03; see Figure 9a).

A Kruskal-Wallis rank sum test found no significant effects on aesthetics ($\chi^2(4)=5.1$, $p=0.277$, r=0.01); see Figure 9b). There was no visualization design to rate in C1-No Vis., so this condition is absent in Figure 9.

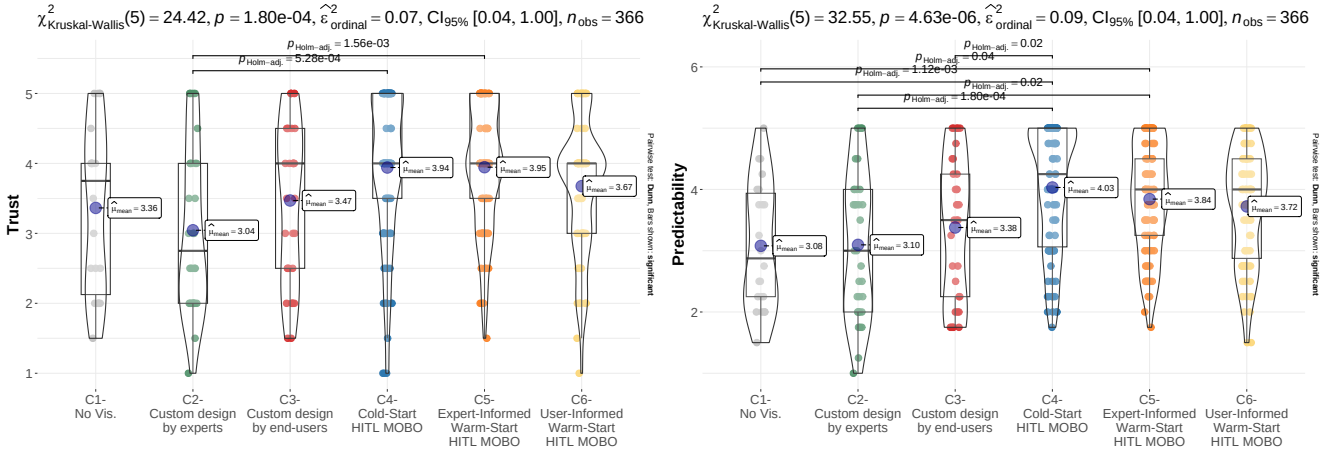
[2]Will be open-sourced – own development.



(a) Statistical comparison of perceived safety.

(b) Statistical comparison of cognitive load.

Figure 7: Statistical comparison of perceived safety and cognitive load over the conditions.



(a) Statistical comparison of trust.

(b) Statistical comparison of predictability.

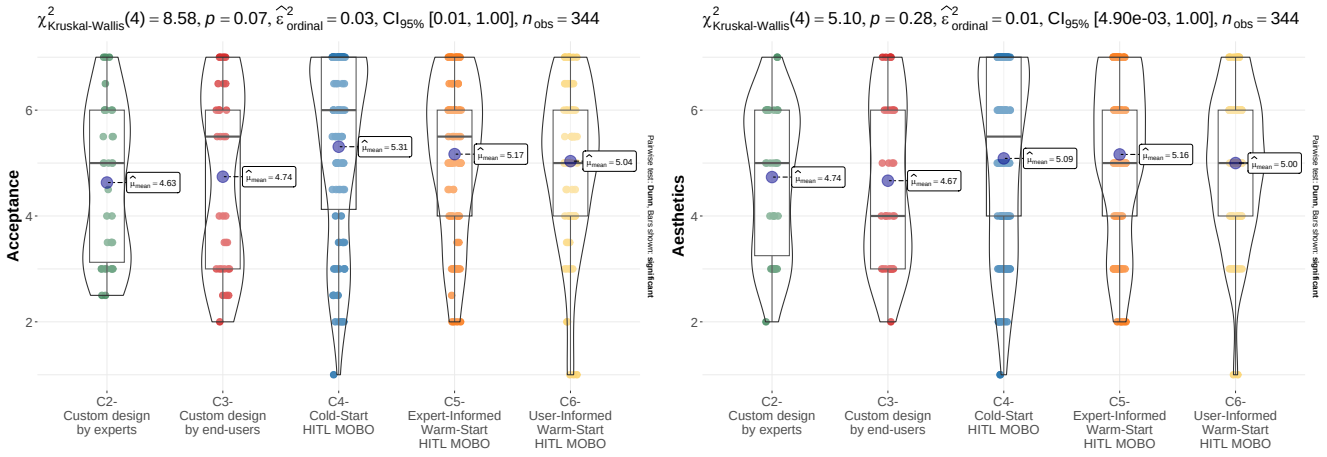
Figure 8: Statistical comparison of trust and predictability over the conditions.

Conclusion. The HITL MOBO conditions (C4-C6) demonstrated significant improvements in *perceived safety*, *trust*, and *predictability* while reducing *cognitive load* compared to the non-MOBO conditions (C2, C3). However, no significant effects for acceptance and aesthetics were observed, providing **partial support for H1**. Moreover, **H2 is rejected**, as no significant differences were identified among the HITL MOBO conditions (C4-C6) for any design objective.

5.1.4 Eye-Tracking Results. The eye-tracking results indicated that participants were attentive to the study, with occasional divergent gazes away from the screen. The ART found a significant main effect of *AOI* ($F(4, 140) = 13.04$, $p<0.001$) and of *visualization condition* on AOI fixation ($F(5, 35) = 14.11$, $p<0.001$). The ART found a significant interaction effect of *AOI* × *visualization condition* on AOI fixation ($F(20, 140) = 2.26$, $p=0.003$; see Figure 15).
For C6-User-Informed Warm-Start HITL MOBO, C2-Custom design by experts, and C3-Custom design by end-users, particular emphasis was placed on the speedometer. In C5-Expert-Informed Warm-Start HITL MOBO, emphasis was put on the car, which was also gazed upon comparatively frequently in the other conditions.

5.2 Pareto Front Parameter Set

Figure 10 shows the final parameter sets per condition (C2–C6), and Figure 11 visualizes these sets within the driving scene. Overall, in the HITL MOBO conditions (C4–C6), the bootstrapped 95% confidence intervals suggest that many parameters were mostly turned off ($v < 0.5$), which contrasts C2 and rarely occurred in C3. Among the HITL MOBO conditions, C4 had the highest proportion



(a) Statistical comparison of acceptance. **(b) Statistical comparison of aesthetics.**

Figure 9: Statistical comparison of acceptance and aesthetics over the conditions. Data on acceptance and aesthetics was not collected for C1-*No Vis*.

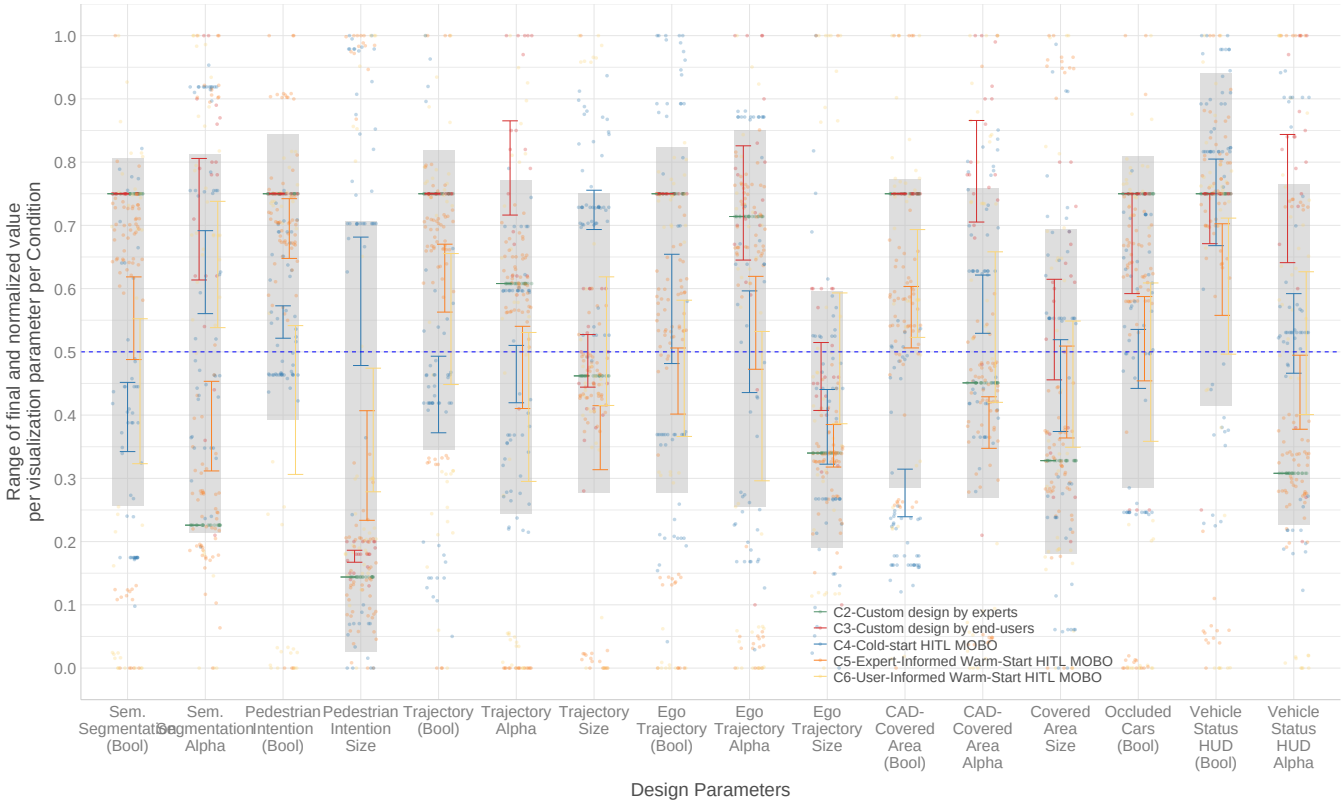


Figure 10: Final parameter set per condition. The jittered Pareto front values per participant are presented, normalized to $[0, 1]$. The gray rectangle shows one standard deviation from the mean of all values. The lines show the bootstrapped 95% confidence intervals per condition. The x-axis shows the ordered parameters (p_1 to p_{16}) from left to right.

of parameters turned off, likely due to the Cold-Start approach not incorporating prior knowledge, for example, informed by the experts' *standard* visualization or users' custom design.

The HITL MOBO conditions exhibit wider intervals than the custom user design condition (C3), indicating a broad range of preferences not captured in the initial custom design phase. C3 showed less variation, possibly because participants found all visualizations initially relevant and enabled them out of curiosity. While these intervals provide insight into commonly preferred parameter ranges, individual differences persist. For instance, some participants activated the *Vehicle Status HUD* while the majority did not, demonstrating that the intervals do not fully capture each user's unique choices.

Regarding visualization transparency, C3 often had intervals above $\alpha = 0.7$, whereas other conditions hovered around $\alpha = 0.5$. This suggests that participants in C3 preferred clearer, more opaque visualizations when first encountering the scenario. Most conditions were similar in size parameters, but the *Pedestrian Intention* visualization in C4–C6 showed larger intervals, pointing to greater preference diversity. This could mean that some participants preferred different parameter values as they became more familiar with the environment. Such variability underscores the value of approaches like OptiCarVis, which can adapt to convergent and divergent user preferences.

5.2.1 User Expectation Conformity, Satisfaction, Confidence, Agency, and Ownership. A Kruskal-Wallis rank sum test found a significant effect of *visualization condition* on user expectation conformity ($\chi^2(4)=10.62$, $p=0.031$, r=0.11). However, post-hoc tests found no significant difference.

Kruskal-Wallis rank sum tests found no significant effects on Satisfaction ($\chi^2(4)=5.21$, $p=0.266$, r=0.06), Confidence ($\chi^2(4)=8.78$, $p=0.067$, r=0.09), Agency ($\chi^2(3)=3.22$, $p=0.358$, r=0.04), or on Ownership ($\chi^2(3)=4.79$, $p=0.188$, r=0.06).

5.3 Qualitative Results

After the final exposure to the visualizations, 22 participants provided open feedback on the expectation, satisfaction, and interactivity themes (see Section 4.5.1). We analyzed this feedback using a structured two-phase process involving three authors. In the first phase, each author independently categorized the feedback into positive, negative, or suggestive sentiments across the three themes, summarizing key statements and insights. In the second phase, the authors collaboratively reviewed and finalized which feedback to include, ensuring consistent sentiment assignment and resolving disagreements.

5.3.1 Expectation. Analyzing the participants' expectations of the design reveals two distinct sentiments. Two positive comments emphasized comfort and safety, suggesting that the absence of excessive information made participants feel safer. One participant stated, *"...for some reason, I felt more comfortable when I was not really seeing much of what the 'car' was 'thinking.'"* Another participant mentioned that the experience was less mentally demanding than expected, implying that an overly complicated interface can cause mental fatigue.

Conversely, negative feedback mainly revolved around design inconsistencies and inadequate visualization. Four participants showed concern when critical visual cues like the blue spheres (CAD-covered area visualization) disappeared, as noted, *"...the blue bubbles showing the coverage area were also gone for some reason."* Others pointed out unnecessary or distracting design elements, such as *"...the red jerky line and three blue lines..."* (referring to the trajectories). The overarching sentiment was a desire for more intuitive visualizations to understand the AV better.

5.3.2 Satisfaction. Satisfaction levels varied among participants. Seven positive comments praised the effective visualization of pedestrians and vehicles (e.g., *"I like the [...] way the car highlights everything, including the pedestrians and other vehicles*". However, four participants voiced concerns about color coordination and the segmentation of certain objects, suggesting that some design elements could potentially confuse or distract the AV user. This sentiment is captured in the statement, *"...some elements were too similar to each other in terms of color..."*.

5.3.3 Interactivity. Interactivity feedback illuminated participants' desire for more control and customization. Six positive remarks highlighted user satisfaction with the design optimization process, suggesting that personalized designs might increase user trust. A participant mentioned, *"I liked having more control over the design. I feel like I would be helping a lot of people."* The seven negative feedbacks highlighted issues with information overload (*"I felt a little overwhelmed when the people, the vehicles, and the signs were all highlighted"*). Seven participants provided suggestions regarding visualization design enhancement and additional visualizations. Common suggestions included the addition of turn signals, clearer indications of the vehicle's route and intentions, and more dash notifications. One participant's comprehensive feedback, *"...I wonder if there could be warning symbols and sounds when cars are braking..."* offers valuable insights into enhancing user trust through proactive system communications.

Conclusion. Participants' feedback indicates increased engagement and alignment with user preferences through the HITL optimization process. The positive and negative feedback underscores the importance of intuitive design, clear visual cues, and user customization in building trust and ensuring user satisfaction.

6 Discussion

6.1 Applicability of Bayesian Optimization on In-Vehicle Visualization Design

The computational approach in OptiCarVis effectively optimized the design objectives, as Cold-Start HITL MOBO (C4) yielded outcomes significantly superior (i.e., perceived safety, trust, predictability, usefulness, and satisfying) to the non-MOBO approaches (C2 and C3), which aligns with HCI literature [9]. These findings partially support H1, as HITL MOBO conditions enhanced user ratings for most objectives but showed no significant improvements in acceptance or aesthetics. This suggests that querying user preferences in detail regarding visual design elements or cultural factors (see [28, 54]) in the HITL optimization process could help capture these subjective aspects more effectively.

The Warm-Start approaches (C5 and C6) also showcased significantly higher ratings than the non-MOBO approaches. However, these were less pronounced. Furthermore, no significant differences were observed between Cold-Start (C4) and Warm-Start (C5 and C6)

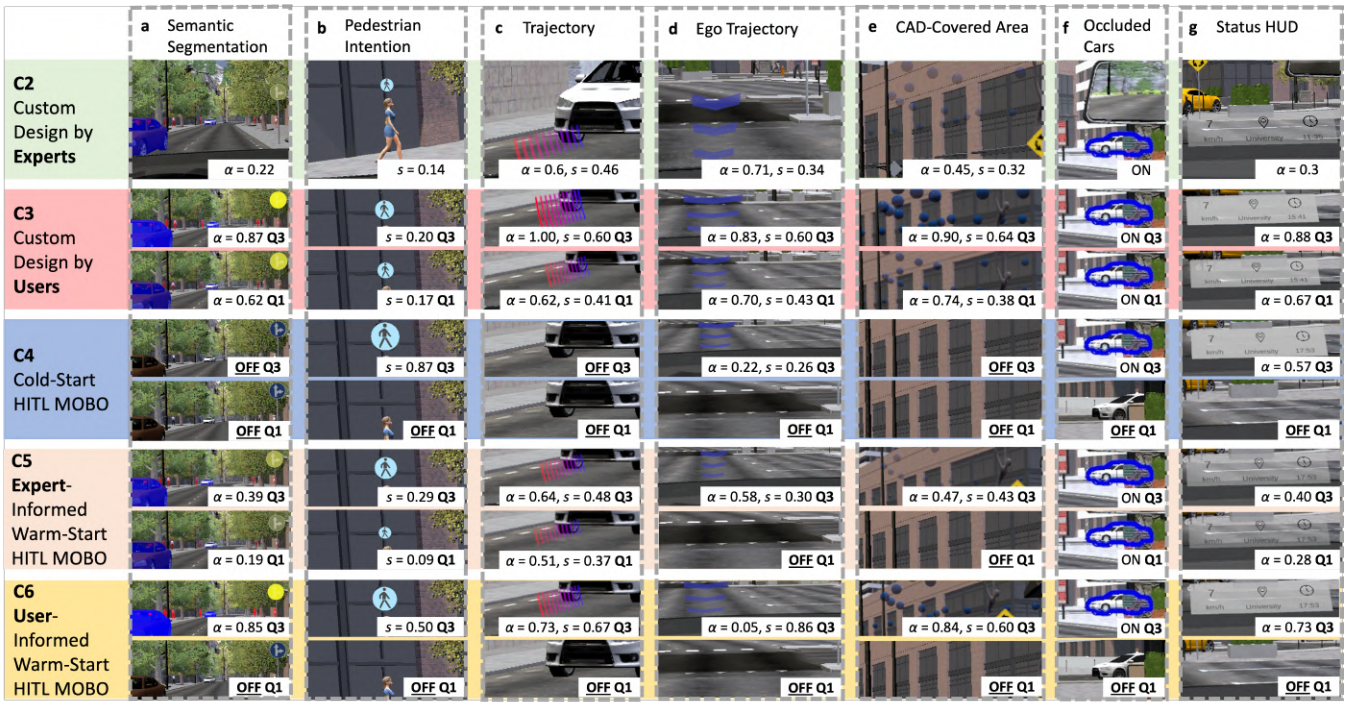


Figure 11: Visualization of parameter values from participants' Pareto front in conditions C3–C6 and the experts' *standard* design in condition C2 (see Figure 10). The parameters displayed are transparency (α), size (s), and visibility (v). A visualization is *OFF* if $v < 0.5$ and *ON* if $v \geq 0.5$. C2 shows the mean parameter values from the experts' design. Conditions C3–C6 present participant parameter values at the interquartile range's lower (25th percentile, Q1) and upper (75th percentile, Q3) ends.

conditions, leading to the rejection of H2, which hypothesized that the user-informed (C6) approach would outperform other HITL conditions. This might be attributed to imperfections in expert or user design data. Expert designs may not scale well to large sample sizes, and novice user custom designs can be flawed because user ratings may change after the initiation of the Bayesian optimizer as users get familiar with the situation. Thus, instead of enhancing the optimizer's exploration of the design space, the effectiveness of the Warm-Start approaches (C5 and C6) was slightly diminished compared to the Cold-Start variant. Also, no significant differences between Cold-Start and Warm-Start HITL MOBO suggest that an averaged expert design could suffice as initial data for the optimizer in the AV functionality visualization design task.

In contrast to visualization designs resulting from traditional approaches (e.g., expert design, see [13, 14, 20, 73, 84]), MOBO can effectively identify more personalized designs. Traditional approaches require larger sample sizes and greater monetary incentives to achieve similar results while evaluating such broad design space in A/B testing. We show that MOBO is (at least) equivalent to traditional visualization design approaches in terms of its applicability by overcoming these testing constraints.

Although some participants reported qualitatively that they felt safer with less excessive information, this does not negate the need for feedback visualizations but requires designs that avoid interface clutter. This underscores a key challenge in AV visualization design, which is to augment information density to increase trust without amplifying the cognitive load [13, 14, 20, 73, 81, 84]. Our findings suggest that the (Cold- and Warm-Start) HITL MOBO adeptly facilitates the optimization of all objectives, obviating the need for any trade-offs.

Safety is a key factor in the automotive sector. While HITL optimization is advantageous, it is critical to ensure that the resulting design alterations do not affect driving safety, if possible, also with future AVs. Although we consciously refrained from consulting some existing standards due to their orientation towards manual rather than automated driving, the OptiCarVis method can be harmonized with such standards. For instance, we constrained certain design parameters, like the position of visualization elements, to avert overlapping UI elements (see Section 3.1). Thus, future work should investigate the balance between customization and standardization in in-vehicle displays, particularly in scenarios where multiple end-users (e.g., in a shared AV) expect varying levels of information and functionality.

Creating scalable solutions for various vehicle models and contexts is challenging. However, our results demonstrate that the OptiCarVis approach effectively generates visualization designs that receive high ratings in *safety*, *trust*, *predictability*, *usefulness*, *satisfaction*, and *aesthetics*. Additionally, as vehicles operate in dynamic environments [6, 42], the performance may improve with detailed driving context information (e.g., see [6]). Thus, regular updates and re-optimizations are crucial, emphasizing the importance

of long-term studies to capture evolving user needs and technology trends.

Besides, our findings indicate a convergence in end-user ratings towards a *satisfying* level that does not necessarily represent the highest possible value of an objective (see Section 3.2). As the optimization progresses, for example, in long-term usage, the potential improvements in user trust, acceptance, or perceived safety from design changes might diminish. This diminishing return can lead to a saturation point, beyond which further optimization might not yield significant benefits or be cost-effective.

6.2 Naturalistic User Reactions to Optimizer-Led Design Processes

Integrating user recordings via webcam-based eye-tracking provides a unique perspective into user engagement and attention patterns during the HITL design process despite the noisiness of the data. For instance, if users know that their design evaluation is being observed and validated, it is more likely that their feedback is genuine (see also Hawthorne effect [66]). This can enhance the validity of the optimization process, ensuring that the ratings queried reflect the users' genuine experiences.

Analyzing the gaze patterns and AoIs revealed varying user attention across different conditions. However, we did not find a uniform pattern of attention across all participants or conditions. Regardless of the MOBO condition, other cars and the speedometer attracted a high fixation percentage. This indicates that such primary driving-related information is important to users. As AV technology becomes more prevalent, it is plausible that users' attention might transition to elements more pertinent to non-driving related activities (see [68]).

Besides, eye-tracking data can be beneficial for refining the optimization process. We might assign a weighted importance metric by ascertaining which UI elements attract the most attention during classification. This adaptive approach could ensure the optimization remains attuned to user preferences and needs.

6.3 Empowering Non-Experts in Designing In-Vehicle Visualizations

Typically, end-users provide feedback during A/B testing of in-vehicle visualization designs as part of a user-centered design process [40]. However, designers must interpret and iteratively integrate this feedback, a process that could benefit from direct incorporation through HITL design to meet individual needs and preferences better. Furthermore, while there is a clear demand for personalizing visualization designs [18], current design approaches often restrict personalization within predefined limits set by designers. Additionally, implementing these manual personalizations can be challenging, frequently requiring users to navigate through setting menus.

OptiCarVis presents a paradigm shift. By leveraging optimization-driven approaches, we demonstrate the potential to empower end-users to participate actively in the in-vehicle visualization design process. This democratizes the design and paves the way for optimized personalization, allowing for designs that simultaneously cater to multiple objectives. Such a framework can bridge the gap between designers' intents and users' preferences. Another potential solution to better harness prior user knowledge could be to introduce a startup phase extending our investigation of Warm-Start HITL MOBO approaches (C5 and C6). This phase could involve querying users about their existing knowledge before their first drive and subsequently through brief follow-up questions. The optimization could be fast with a broad user base, and Pareto's optimal designs could be approached quickly.

However, limitations exist, such as the explicit nature of the HITL process with subjective feedback and the iterative nature of multiple optimization iterations [9, 57].

6.4 Towards Implicit Design Optimizations of In-Vehicle Visualizations

Our MOBO approach employs an explicit optimization loop that continuously requests feedback on users' subjective states. However, frequent queries can lead to user fatigue and reduce the accuracy of their responses [9]. Additionally, these requests may disrupt users' ongoing in-vehicle UI interactions. Taking inspiration from Koyama and Goto [49], a shift from an explicit to an implicit optimization loop is conceivable. They leverage BO to learn design objectives by observing design exploration behaviors. The system then offers design suggestions based on these observations. Likewise, we envisage an implicit MOBO process incorporated into vehicle use. Although this system may still operate within a loop, it would transition from explicit feedback (e.g., Likert scale ratings) to implicit end-user feedback. Such feedback could be derived from their interaction behaviors like the input error rate, physiological cues like heart rate, or psychological indicators like emotional states. Relevant approaches in this domain were discussed by Stampf et al. [78] and Colley, Hartwig et al. [15], emphasizing non-intrusive feedback collection during vehicle use via interior cameras.

6.5 Necessity of Optimization Explainability

The explainability of automated systems, particularly in the context of feedback visualizations in AVs, is pivotal [13, 14, 20, 47, 73, 84]. Trust in automated systems, an essential component for user acceptance, is often intertwined with the user's comprehension of the system's behavior [13, 14, 20]. Thus, users may require a sound understanding of the mechanisms underpinning the adaptive nature of the UI enabled by OptiCarVis to foster trust.

However, the challenge lies in communicating complex optimization algorithms, like MOBO, both transparent and comprehensible to non-expert users. Our study revealed that user agency and satisfaction remained high across all conditions (MOBO and non-MOBO). This suggests that participants also were content with the overall optimization process. Still, it is unclear to what extent they have understood it. It is noteworthy, though, to discern which specific facets of the optimization process contributed to the sense of satisfaction and perceived agency. Our results align with and extend the findings by Chan et al. [9]. Future research endeavors should delve deeper into understanding the nuances of users and how they shape their interactions with automated systems.

6.6 Limitations and Future Work

Our work has been instrumental in assessing the application of MOBO to improve the user experience of AV functionality visualizations. However, limitations exist. The choice of algorithm, including the acquisition function and other parameters, might have influenced the study outcomes. As with any optimization approach, the selected hyperparameters can significantly impact the results. Along with this, the participant selection might introduce biases, which could affect the generalizability of our findings. Besides, a limitation was using a webcam-based eye-tracking method, which has inherent inaccuracies [85]. Nonetheless, we argue that the inaccuracies are a necessary trade-off, primarily because the technique offered novel insights into users' reactions to the HITL process and the resulting designs that are otherwise infeasible for an online study.

The study's 33-second observation period per MOBO iteration is brief, which may limit the depth of understanding about the AV's functionalities and limits. However, previous work also employed short durations (e.g., one minute [20]). Another limitation is that the 33-second route cannot cover every possible driving scenario. However, the optimal parameter values likely vary depending on the scenario. We argue that our results can still be generalized to most urban scenarios with cars, pedestrians, road crossings, and roundabouts. This is also shown by the final 3-minute route, where the optimized designs were still effective even though the scenario changed (e.g., an intersection was introduced). However, future work should investigate other environments, such as motorways, and consider different traffic dynamics, such as pedestrian densities.

In the MOBO conditions, participants had more exposure to the 33-second route than those in non-MOBO conditions. This repeated exposure, an inherent aspect in iterative HITL processes, likely increased familiarity with the scenario. However, we argue this was mitigated by showing the unknown 3-minute route for the final assessment across all conditions and the frequently changing visualization designs during MOBO iterations. Yet, future work should explore increased scenario exposure in non-MOBO conditions to investigate this further.

While the online study environment with driving videos on a computer screen provides a controlled environment with high internal validity, it may have implications for generalizability in real-world driving scenarios. Our results are generalizable despite the wider field of view and possible physical consequences in the real world. In the real world, end-users would also focus on the scene directly in front of the AV and be less interested in what happens behind it. Even without real-world consequences, participants can realistically reflect their subjective perceptions in driving simulations [83]. Besides, considering the complexity of real-world driving scenarios and the potential traffic dangers our prototype system could have introduced, a simulated environment was deemed suitable for this initial exploration of HITL MOBO in the automotive domain. Yet, future studies should prioritize the execution in real vehicles under dynamic driving conditions, for example, using approaches like XR-OOM [34] or PassengXR [63]. If technically impractical, a vehicle motion simulator can be used (e.g., SwiVR Car-Seat [16]).

Interface clutter is a recognized issue [13, 14, 20]. An interface attempting to display all potential visualization parameters could become overwhelming, potentially distracting users or obscuring essential information. In our study, while we aimed for comprehensive visualization, the risk of cluttering the interface remained evident.

Cultural differences influence user perceptions of AVs [28, 54], for instance, the acceptance of AV maneuvers [28], also likely affecting their perception of AV functionality visualizations. Our study was limited to one culture. However, the OptiCarVis approach using HITL optimization could accommodate diverse user needs stemming from cultural context. To assess its effectiveness, future research should evaluate OptiCarVis with users from different cultures.

7 Conclusion

This work employed HITL MOBO to navigate the design space of AV functionality visualizations on AR WSDs. An online study with N=117 participants helped validate the MOBO-driven approach, confirming its efficacy in optimizing visualizations for AVs. While statistical significance was not reached compared to No Visualization (C1) besides for predictability, the ratings were **always** better for C4-Cold-Start HITL MOBO. Besides, this Cold-Start and Warm-Start optimization significantly improved perceived safety, cognitive load, and trust. Acceptance and aesthetics did not differ over the conditions. Consequently, OptiCarVis provides a pathway for creating personalized in-vehicle visualization designs that improve end-user experiences and decrease development time and expenses.

Open Science

We make the Bayesian optimizer (see https://github.com/Pascal-Jansen/Bayesian-Optimization-for-Unity), the Unity application upon request, and the collected (anonymized) data (see https://github.com/M-Colley/opticarvis-data) available. The Unity project supports novel application scenarios by including easily adaptable settings regarding MOBO, server-client infrastructure for the online study, webcam-based eye tracking [82], and AV driving behavior.

Acknowledgments

We thank all study participants. Additionally, we thank the 6th Summer School on Computational Interaction[3] for providing significant insights into computational optimization. This work was supported by a German Academic Exchange Service (DAAD) fellowship.

References

[1] Nadia Adnan, Shahrina Md Nordin, Mohamad Ariff bin Bahruddin, and Murad Ali. 2018. How trust can drive forward the user acceptance to the technology? In-vehicle technology for autonomous vehicle. *Transportation research part A: policy and practice* 118 (2018), 819–836.

[2] Stéphane Alarie, Charles Audet, Aïmen E Gheribi, Michael Kokkolaras, and Sébastien Le Digabel. 2021. Two decades of blackbox optimization applications. *EURO Journal on Computational Optimization* 9 (2021), 100011.

[3] Jackie Ayoub, Feng Zhou, Shan Bao, and X. Jessie Yang. 2019. From Manual Driving to Automated Driving: A Review of 10 Years of AutoUI. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Utrecht, Netherlands) *(AutomotiveUI '19)*. ACM, New York, NY, USA, 70–90. doi:10.1145/3342197.3344529

[3] https://cixschool2022.cs.uni-saarland.de/

[4] Maximilian Balandat, Brian Karrer, Daniel R. Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. 2020. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. In *Advances in Neural Information Processing Systems 33*. http://arxiv.org/abs/1910.06403

[5] Johannes Beller, Matthias Heesen, and Mark Vollrath. 2013. Improving the Driver–Automation Interaction: An Approach Using Automation Uncertainty. *Human Factors* 55, 6 (2013), 1130–1141. doi:10.1177/0018720813482327 arXiv:https://doi.org/10.1177/0018720813482327 PMID: 24745204.

[6] David Bethge, Thomas Kosch, Tobias Grosse-Puppendahl, Lewis L. Chuang, Mohamed Kari, Alexander Jagaciak, and Albrecht Schmidt. 2021. VEmotion: Using Driving Context for Indirect Emotion Prediction in Real-Time. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. ACM, New York, NY, USA, 638–651. doi:10.1145/3472749.3474775

[7] Ali Borji and Laurent Itti. 2013. Bayesian optimization explains human active search. *Advances in neural information processing systems* 26 (2013).

[8] Eric Brochu, Vlad M Cora, and Nando De Freitas. 2010. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599* (2010).

[9] Liwei Chan, Yi-Chi Liao, George B Mo, John J Dudley, Chun-Lien Cheng, Per Ola Kristensson, and Antti Oulasvirta. 2022. Investigating Positive and Negative Qualities of Human-in-the-Loop Optimization for Designing Interaction Techniques. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. ACM, New York, NY, USA, Article 112, 14 pages. doi:10.1145/3491102.3501850

[10] Suyog Chandramouli, Yifan Zhu, and Antti Oulasvirta. 2023. Interactive Personalization of Classifiers for Explainability Using Multi-Objective Bayesian Optimization. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization* (Limassol, Cyprus) *(UMAP '23)*. ACM, New York, NY, USA, 34–45. doi:10.1145/3565472.3592956

[11] Chia-Hsing Chiu, Yuki Koyama, Yu-Chi Lai, Takeo Igarashi, and Yonghao Yue. 2020. Human-in-the-Loop Differential Subspace Search in High-Dimensional Latent Space. *ACM Trans. Graph.* 39, 4, Article 85 (aug 2020), 15 pages. doi:10.1145/3386569.3392409

[12] Toby Chong, I-Chao Shen, Issei Sato, and Takeo Igarashi. 2021. Interactive Optimization of Generative Image Modelling using Sequential Subspace Search and Content-based Guidance. In *Computer Graphics Forum*, Vol. 40. Wiley Online Library, 279–292.

[13] Mark Colley, Christian Bräuner, Mirjam Lanzer, Marcel Walch, Martin Baumann, and Enrico Rukzio. 2020. Effect of Visualization of Pedestrian Intention Recognition on Trust and Cognitive Load. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Virtual Event, DC, USA) *(AutomotiveUI '20)*. ACM, New York, NY, USA, 181–191. doi:10.1145/3409120.3410648

[14] Mark Colley, Benjamin Eder, Jan Ole Rixen, and Enrico Rukzio. 2021. Effects of Semantic Segmentation Visualization on Trust, Situation Awareness, and Cognitive Load in Highly Automated Vehicles. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, Article 155, 11 pages. https://doi.org/10.1145/3411764.3445351

[15] Mark Colley, Sebastian Hartwig, Albin Zeqiri, Timo Ropinski, and Enrico Rukzio. 2024. AutoTherm: A Dataset and Benchmark for Thermal Comfort Estimation Indoors and in Vehicles. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (Sept. 2024), 49. doi:10.1145/3678503

[16] Mark Colley, Pascal Jansen, Enrico Rukzio, and Jan Gugenheimer. 2022. SwiVR-Car-Seat: Exploring Vehicle Motion Effects on Interaction Quality in Virtual Reality Automated Driving Using a Motorized Swivel Seat. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 150 (dec 2022), 26 pages. doi:10.1145/3494968

[17] Mark Colley, Svenja Krauss, Mirjam Lanzer, and Enrico Rukzio. 2021. How Should Automated Vehicles Communicate Critical Situations? A Comparative Analysis of Visualization Concepts. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3, Article 94 (sep 2021), 23 pages. doi:10.1145/3478111

[18] Mark Colley, Mirjam Lanzer, Jan Henry Belz, Marcel Walch, and Enrico Rukzio. 2021. Evaluating the Impact of Decals on Driver Stereotype Perception and Exploration of Personalization of Automated Vehicles via Digital Decals. In *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Leeds, United Kingdom) *(AutomotiveUI '21)*. ACM, New York, NY, USA, 296–306. doi:10.1145/3409118.3475132

[19] Mark Colley, Luca-Maxim Meinhardt, Alexander Fassbender, Michael Rietzler, and Enrico Rukzio. 2023. Come Fly With Me: Investigating the Effects of Path Visualizations in Automated Urban Air Mobility. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 2, Article 52 (jun 2023), 23 pages. doi:10.1145/3596249

[20] 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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 49 (jul 2022), 21 pages. doi:10.1145/3534609

[21] Mark Colley, Oliver Speidel, Jan Strohbeck, Jan Ole Rixen, Jan Henry Belz, and Enrico Rukzio. 2024. Effects of Uncertain Trajectory Prediction Visualization in Highly Automated Vehicles on Trust, Situation Awareness, and Cognitive Load. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 4, Article 153 (Jan. 2024), 23 pages. doi:10.1145/3631408

[22] Rebecca Currano, So Yeon Park, Dylan James Moore, Kent Lyons, and David Sirkin. 2021. Little Road Driving HUD: Heads-Up Display Complexity Influences Drivers' Perceptions of Automated Vehicles. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, Article 511, 15 pages. https://doi.org/10.1145/3411764.3445575

[23] Fred D Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly* (1989), 319–340.

[24] Klaus Dietmayer. 2016. Predicting of machine perception for automated driving. *Autonomous driving: technical, legal and social aspects* (2016), 407–424.

[25] Tim Driesen-Micklitz, Michael Fellmann, and Carsten Röcker. 2023. A Set of Design Principles for Personalized Information in Automated Driving User Interfaces Based on Theory and Empirical Evidence. In *2023 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, New York, NY, USA, 1–6. doi:10.1109/IV55152.2023.10186755

[26] John J. Dudley, Jason T. Jacques, and Per Ola Kristensson. 2019. Crowdsourcing Interface Feature Design with Bayesian Optimization. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, 1–12. doi:10.1145/3290605.3300482

[27] Mark Dunlop and John Levine. 2012. Multidimensional Pareto Optimization of Touchscreen Keyboards for Speed, Familiarity and Improved Spell Checking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. ACM, New York, NY, USA, 2669–2678. doi:10.1145/2207676.2208659

[28] Aaron Edelmann, Stefan Stümper, and Tibor Petzoldt. 2021. Cross-cultural differences in the acceptance of decisions of automated vehicles. *Applied ergonomics* 92 (2021), 103346.

[29] Stefanie M. Faas, Andrea C. Kao, and Martin Baumann. 2020. A Longitudinal Video Study on Communicating Status and Intent for Self-Driving Vehicle – Pedestrian Interaction. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. ACM, New York, NY, USA, 1–14. doi:10.1145/3313831.3376484

[30] Daniel J Fagnant and Kara Kockelman. 2015. Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy and Practice* 77 (2015), 167–181.

[31] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods* 41, 4 (2009), 1149–1160.

[32] Lukas A. Flohr, Joseph Sebastian Valiyaveettil, Antonio Krüger, and Dieter P. Wallach. 2023. Prototyping Autonomous Vehicle Windshields with AR and Real-Time Object Detection Visualization: An On-Road Wizard-of-Oz Study. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23)*. ACM, New York, NY, USA, 2123–2137. doi:10.1145/3563657.3596051

[33] M. Ghazizadeh, J.D. Lee, and L.N. Boyle. 2012. Extending the Technology Acceptance Model to assess automation. *Cogn Tech Work* 14 (2012), 39–49. https://doi.org/10.1007/s10111-011-0194-3

[34] David Goedicke, Alexandra W.D. Bremers, Sam Lee, Fanjun Bu, Hiroshi Yasuda, and Wendy Ju. 2022. XR-OOM: MiXed Reality Driving Simulation with Real Cars for Research and Design. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. ACM, New York, NY, USA, Article 107, 13 pages. doi:10.1145/3491102.3517704

[35] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, Amsterdam, The Netherlands, 139–183.

[36] Renate Häuslschmid, Max von Bülow, Bastian Pfleging, and Andreas Butz. 2017. Supporting Trust in Autonomous Driving. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces* (Limassol, Cyprus) *(IUI '17)*. ACM, New York, NY, USA, 319–329. doi:10.1145/3025171.3025198

[37] Vincent Hayward, Jehangir Choksi, Gonzalo Lanvin, and Christophe Ramstein. 1994. Design and multi-objective optimization of a linkage for a haptic interface. *Advances in robot kinematics and computational geometry* (1994), 359–368.

[38] Tove Helldin, Göran Falkman, Maria Riveiro, and Staffan Davidsson. 2013. Presenting System Uncertainty in Automotive UIs for Supporting Trust Calibration in Autonomous Driving. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Eindhoven, Netherlands) *(AutomotiveUI '13)*. ACM, New York, NY, USA, 210–217. doi:10.1145/2516540.2516554

[39] International Organization for Standardization. 2017. Road vehicles – Ergonomic aspects of transport information and control systems – Dialogue management principles and compliance procedures. ISO Standard. https://www.iso.org/standard/69238.html Accessed: 2024-02-21.

[40] International Organization for Standardization. 2018. Ergonomics of human-system interaction – Part 11: Usability: Definitions and concepts. ISO Standard. https://www.iso.org/standard/63500.html Accessed: 2024-02-21.

[41] International Organization for Standardization. 2019. Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems.

ISO Standard. https://www.iso.org/standard/77520.html Accessed: 2024-02-21.
[42] Pascal Jansen, Julian Britten, Alexander Häusele, Thilo Segschneider, Mark Colley, and Enrico Rukzio. 2023. AutoVis: Enabling Mixed-Immersive Analysis of Automotive User Interface Interaction Studies. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. ACM, New York, NY, USA, Article 378, 23 pages. doi:10.1145/3544548.3580760
[43] Pascal Jansen, Mark Colley, Tim Pfeifer, and Enrico Rukzio. 2024. Visualizing imperfect situation detection and prediction in automated vehicles: Understanding users' perceptions via user-chosen scenarios. *Transportation Research Part F: Traffic Psychology and Behaviour* 104 (2024), 88–108.
[44] Pascal Jansen, Mark Colley, and Enrico Rukzio. 2022. A Design Space for Human Sensor and Actuator Focused In-Vehicle Interaction Based on a Systematic Literature Review. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 56 (jul 2022), 51 pages. doi:10.1145/3534617
[45] Florian Kadner, Yannik Keller, and Constantin Rothkopf. 2021. AdaptiFont: Increasing Individuals' Reading Speed with a Generative Font Model and Bayesian Optimization. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. ACM, New York, NY, USA, Article 585, 11 pages. doi:10.1145/3411764.3445140
[46] Mohammad M. Khajah, Brett D. Roads, Robert V. Lindsey, Yun-En Liu, and Michael C. Mozer. 2016. Designing Engaging Games Using Bayesian Optimization. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 5571–5582. doi:10.1145/2858036.2858253
[47] Jeamin Koo, Jungsuk Kwac, Wendy Ju, Martin Steinert, Larry Leifer, and Clifford Nass. 2015. Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *International Journal on Interactive Design and Manufacturing (IJIDeM)* 9, 4 (2015), 269–275.
[48] Moritz Körber. 2019. Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*, Sebastiano Bagnara, Riccardo Tartaglia, Sara Albolino, Thomas Alexander, and Yushi Fujita (Eds.). Springer International Publishing, Cham, 13–30.
[49] Yuki Koyama and Masataka Goto. 2022. BO as Assistant: Using Bayesian Optimization for Asynchronously Generating Design Suggestions. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. ACM, New York, NY, USA, Article 77, 14 pages. doi:10.1145/3526113.3545664
[50] Yuki Koyama, Issei Sato, and Masataka Goto. 2020. Sequential Gallery for Interactive Visual Design Optimization. *ACM Trans. Graph.* 39, 4, Article 88 (aug 2020), 12 pages. doi:10.1145/3386569.3392444
[51] Alexander Kunze, Stephen J. Summerskill, Russell Marshall, and Ashleigh J. Filtness. 2018. Augmented Reality Displays for Communicating Uncertainty Information in Automated Driving. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Toronto, ON, Canada) *(AutomotiveUI '18)*. ACM, New York, NY, USA, 164–175. doi:10.1145/3239060.3239074
[52] Alexander Kunze, Stephen J. Summerskill, Russell Marshall, and Ashleigh J. Filtness. 2019. Conveying Uncertainties Using Peripheral Awareness Displays in the Context of Automated Driving. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Utrecht, Netherlands) *(AutomotiveUI '19)*. ACM, New York, NY, USA, 329–341. doi:10.1145/3342197.3344537
[53] Miltos Kyriakidis, Riender Happee, and Joost CF de Winter. 2015. Public opinion on automated driving: Results of an international questionnaire among 5000 respondents. *Transportation research part F: traffic psychology and behaviour* 32 (2015), 127–140.
[54] Mirjam Lanzer, Franziska Babel, Fei Yan, Bihan Zhang, Fang You, Jianmin Wang, and Martin Baumann. 2020. Designing Communication Strategies of Autonomous Vehicles with Pedestrians: An Intercultural Study. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, New York, NY, USA, 122–131.
[55] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.
[56] Yi-Chi Liao. 2023. Human-in-the-Loop Design Optimization. (2023).
[57] Yi-Chi Liao, John J Dudley, George B Mo, Chun-Lien Cheng, Liwei Chan, Antti Oulasvirta, and Per Ola Kristensson. 2023. Interaction Design With Multi-objective Bayesian Optimization. *IEEE Pervasive Computing* 22, 1 (2023), 29–38.
[58] Yi-Chi Liao, George B Mo, John J Dudley, Chun-Lien Cheng, Liwei Chan, Per Ola Kristensson, and Antti Oulasvirta. 2024. Practical approaches to group-level multi-objective Bayesian optimization in interaction technique design. *Collective Intelligence* 3, 1 (2024), 26339137241241313.
[59] Patrick Lindemann, Tae-Young Lee, and Gerhard Rigoll. 2018. Catch my drift: Elevating situation awareness for highly automated driving with an explanatory windshield display user interface. *Multimodal Technologies and Interaction* 2, 4 (2018), 71.
[60] Martina Mara and Kathrin Meyer. 2022. Acceptance of autonomous vehicles: An overview of user-specific, car-specific and contextual determinants. *User experience design in the era of automated driving* (2022), 51–83.
[61] R.T. Marler and J.S. Arora. 2004. Survey of multi-objective optimization methods for engineering. *Structural and Multidisciplinary Optimization* 26, 6 (April 2004), 369–395. doi:10.1007/s00158-003-0368-6
[62] David Issa Mattos, Jan Bosch, Helena Holmstrom Olsson, Aita Maryam Korshani, and Jonn Lantz. 2020. Automotive A/B testing: Challenges and lessons learned from practice. In *2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*. IEEE, New York, NY, USA, 101–109.
[63] Mark McGill, Graham Wilson, Daniel Medeiros, and Stephen Anthony Brewster. 2022. PassengXR: A Low Cost Platform for Any-Car, Multi-User, Motion-Based Passenger XR Experiences. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. ACM, New York, NY, USA, Article 2, 15 pages. doi:10.1145/3526113.3545657
[64] Tobias Müller, Mark Colley, Gülsemin Dogru, and Enrico Rukzio. 2022. AR4CAD: Creation and Exploration of a Taxonomy of Augmented Reality Visualization for Connected Automated Driving. *Proc. ACM Hum.-Comput. Interact.* 6, MHCI, Article 177 (sep 2022), 27 pages. doi:10.1145/3546712
[65] Carl Jörgen Normark. 2015. Design and Evaluation of a Touch-Based Personalizable In-Vehicle User Interface. *International Journal of Human–Computer Interaction* 31, 11 (2015), 731–745. doi:10.1080/10447318.2015.1045240 arXiv:https://doi.org/10.1080/10447318.2015.1045240
[66] David Oswald, Fred Sherratt, and Simon Smith. 2014. Handling the Hawthorne effect: The challenges surrounding a participant observer. *Review of social studies* 1, 1 (2014), 53–73.
[67] Changkun Ou, Daniel Buschek, Sven Mayer, and Andreas Butz. 2022. The Human in the Infinite Loop: A Case Study on Revealing and Explaining Human-AI Interaction Loop Failures. In *Proceedings of Mensch Und Computer 2022* (Darmstadt, Germany) *(MuC '22)*. ACM, New York, NY, USA, 158–168. doi:10.1145/3543758.3543761
[68] Bastian Pfleging, Maurice Rang, and Nora Broy. 2016. Investigating User Needs for Non-Driving-Related Activities during Automated Driving. In *Proceedings of the 15th International Conference on Mobile and Ubiquitous Multimedia* (Rovaniemi, Finland) *(MUM '16)*. ACM, New York, NY, USA, 91–99. doi:10.1145/3012709.3012735
[69] Matthias Poloczek, Jialei Wang, and Peter I Frazier. 2016. Warm starting Bayesian optimization. In *2016 Winter simulation conference (WSC)*. IEEE, IEEE, New York, NY, USA, 770–781.
[70] Amir Rasouli and John K Tsotsos. 2019. Autonomous vehicles that interact with pedestrians: A survey of theory and practice. *IEEE Transactions on Intelligent Transportation Systems* 21, 3 (2019), 900–918.
[71] Andreas Riegler, Andreas Riener, and Clemens Holzmann. 2019. Adaptive dark mode: Investigating text and transparency of windshield display content for automated driving. *Mensch und Computer 2019-Workshopband* (2019).
[72] SAE International. [n. d.]. SAE Levels of Driving Automation™ Refined for Clarity and International Audience. https://www.sae.org/blog/sae-j3016-update.
[73] Tobias Schneider, Joana Hois, Alischa Rosenstein, Sabiha Ghellal, Dimitra Theofanou-Fülbier, and Ansgar R.S. Gerlicher. 2021. ExplAIn Yourself! Transparency for Positive UX in Autonomous Driving. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, Article 161, 12 pages. https://doi.org/10.1145/3411764.3446647
[74] Brandon Schoettle and Michael Sivak. 2014. *A survey of public opinion about autonomous and self-driving vehicles in the US, the UK, and Australia*. Technical Report. University of Michigan, Ann Arbor, Transportation Research Institute.
[75] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. 2015. Taking the human out of the loop: A review of Bayesian optimization. *Proc. IEEE* 104, 1 (2015), 148–175.
[76] James Shanteau, David J Weiss, Rickey P Thomas, Julia Pounds, and Bluemont Hall. 2003. How can you tell if someone is an expert? Empirical assessment of expertise. *Emerging perspectives on judgment and decision research* (2003), 620–641.
[77] Srinath Sridhar, Anna Maria Feit, Christian Theobalt, and Antti Oulasvirta. 2015. Investigating the Dexterity of Multi-Finger Input for Mid-Air Text Entry. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. ACM, New York, NY, USA, 3643–3652. doi:10.1145/2702123.2702136
[78] Annika Stampf, Mark Colley, and Enrico Rukzio. 2022. Towards Implicit Interaction in Highly Automated Vehicles - A Systematic Literature Review. *Proc. ACM Hum.-Comput. Interact.* 6, MHCI, Article 191 (sep 2022), 21 pages. doi:10.1145/3546726
[79] H. Takagi. 2001. Interactive evolutionary computation: fusion of the capabilities of EC optimization and human evaluation. *Proc. IEEE* 89, 9 (2001), 1275–1296. doi:10.1109/5.949485
[80] Jinke D Van Der Laan, Adriaan Heino, and Dick De Waard. 1997. A simple procedure for the assessment of acceptance of advanced transport telematics. *Transportation Research Part C: Emerging Technologies* 5, 1 (1997), 1–10.
[81] Tamara von Sawitzky, Philipp Wintersberger, Andreas Riener, and Joseph L. Gabbard. 2019. Increasing Trust in Fully Automated Driving: Route Indication on an Augmented Reality Head-up Display. In *Proceedings of the 8th ACM International*

Symposium on Pervasive Displays (Palermo, Italy) *(PerDis '19)*. ACM, New York, NY, USA, Article 6, 7 pages. doi:10.1145/3321335.3324947

[82] Tobias Wagner, Mark Colley, Daniel Breckel, Michael Kösel, and Enrico Rukzio. 2024. UnitEye: Introducing a User-Friendly Plugin to Democratize Eye Tracking Technology in Unity Environments. In *Proceedings of Mensch Und Computer 2024* (Karlsruhe, Germany) *(MuC '24)*. ACM, New York, NY, USA, 1–10. doi:10.1145/3670653.3670655

[83] Ying Wang, Bruce Mehler, Bryan Reimer, Vincent Lammers, Lisa A D'Ambrosio, and Joseph F Coughlin. 2010. The validity of driving simulation for assessing differences between in-vehicle informational interfaces: A comparison with field testing. *Ergonomics* 53, 3 (2010), 404–420.

[84] Scott R Winter, Stephen Rice, Nadine K Ragbir, Bradley S Baugh, Mattie N Milner, Bee-Ling Lim, John Capps, and E Anania. 2019. *Assessing pedestrians' perceptions and willingness to interact with autonomous vehicles.* Technical Report. US Department of Transportation. Center for Advanced Transportation Mobility ....

[85] Katarzyna Wisiecka, Krzysztof Krejtz, Izabela Krejtz, Damian Sromek, Adam Cellary, Beata Lewandowska, and Andrew Duchowski. 2022. Comparison of Webcam and Remote Eye Tracking. In *2022 Symposium on Eye Tracking Research and Applications* (Seattle, WA, USA) *(ETRA '22)*. ACM, New York, NY, USA, Article 32, 7 pages. doi:10.1145/3517031.3529615

[86] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. ACM, New York, NY, USA, 143–146. doi:10.1145/1978942.1978963

[87] Yukinobu Nakamura. 2008. *JAMA Guideline for In-Vehicle Display Systems.* SAE Technical Paper 2008-21-0003. SAE International. https://www.sae.org/publications/technical-papers/content/2008-21-0003/ Accessed on 2024-02-28.

[88] Cheng Yunuo, Zhong Xia, Ye Min, and Tian Liwei. 2022. Usability Evaluation of in-Vehicle AR-HUD Interface Applying AHP-GRA. *Human-Centric Intelligent Systems* 2, 3-4 (2022), 124–137.

[89] Qiaoning Zhang, Connor Esterwood, Anuj K. Pradhan, Dawn Tilbury, X. Jessie Yang, and Lionel P. Robert. 2023. The Impact of Modality, Technology Suspicion, and NDRT Engagement on the Effectiveness of AV Explanations. *IEEE Access* 11 (2023), 81981–81994. doi:10.1109/ACCESS.2023.3302261

[90] Mingyuan Zhong, Gang Li, and Yang Li. 2021. Spacewalker: Rapid UI Design Exploration Using Lightweight Markup Enhancement and Crowd Genetic Programming. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. ACM, New York, NY, USA, Article 315, 11 pages. doi:10.1145/3411764.3445326

[91] Xia Zhong, Yunuo Cheng, Jiahao Yang, and Liwei Tian. 2023. Evaluation and Optimization of In-Vehicle HUD Design by Applying an Entropy Weight-VIKOR Hybrid Method. *Applied Sciences* 13, 6 (2023). doi:10.3390/app13063789

A Expert Study to Inform the Standard Visualization Design

B Procedure

AV functionality introduction:

> *You will see a video of a driving session in an automated vehicle. The vehicle takes over lateral and longitudinal control (braking, accelerating, steering). The vehicle attempts to assess the situation and determine the intent of nearby pedestrians and cars. While watching the video, you are supposed to imagine sitting in such an automated vehicle, following the entire journey attentively, and then assessing it.*

Introduction to the trips for C4-C6.

> *After each trip, please rate the visualization. The vehicle will adapt its design based on your feedback. This continues until the vehicle finalizes a design, and you'll take one more ride with that selected design.*

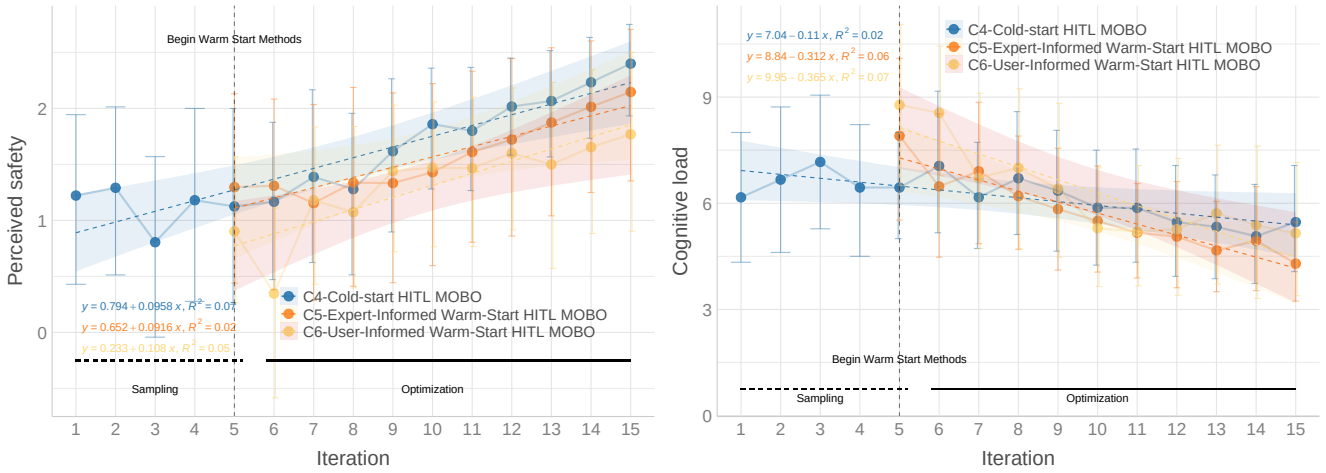
C Results

C.1 Objective Values over Iterations

C.2 Eye-Tracking Results

Table 2: Results of the expert study regarding the 16 design parameters normalized to a $[0, 1]$ range. Experts (E1 - E8) used the custom parameter design tool to manually select parameters they deemed fitting for the given driving scenario.

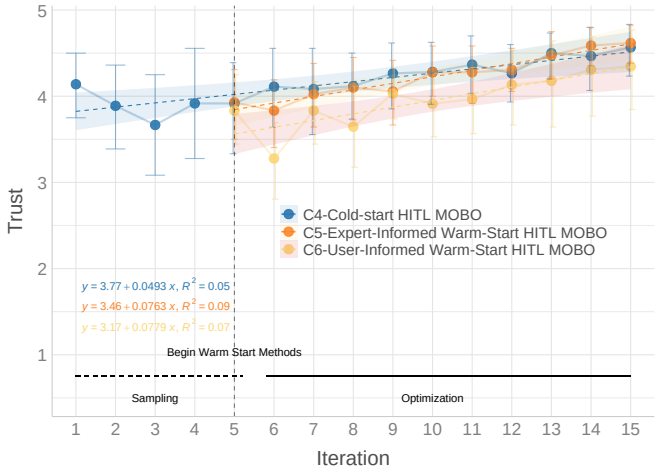
Design Parameter	E1	E2	E3	E4	E5	E6	E7	E8	Mean	SD
p_1: Sem. Segmentation	0.75	0.75	0.75	0.75	0.75	0.75	0.25	0.75	0.69	0.18
p_2: Sem. Segmentation Alpha	0.10	0.37	0.10	0.10	0.19	0.10	0.36	0.49	0.23	0.16
p_3: Pedestrian Intention	0.75	0.75	0.25	0.75	0.75	0.75	0.75	0.75	0.69	0.18
p_4: Pedestrian Intention Size	0.19	0.20	0.14	0.10	0.11	0.15	0.10	0.17	0.15	0.04
p_5: Trajectory	0.75	0.75	0.75	0.75	0.75	0.25	0.75	0.75	0.69	0.18
p_6: Trajectory Alpha	0.63	1.00	0.45	0.41	0.31	0.54	0.52	1.00	0.61	0.26
p_7: Trajectory Size	0.52	0.59	0.60	0.51	0.17	0.34	0.46	0.51	0.46	0.14
p_8: Ego Trajectory	0.75	0.25	0.75	0.75	0.75	0.25	0.25	0.75	0.56	0.26
p_9: Ego Trajectory Alpha	0.72	0.87	0.71	0.53	0.76	0.55	0.97	0.60	0.71	0.15
p_{10}: Ego Trajectory Size	0.41	0.52	0.17	0.28	0.26	0.35	0.35	0.38	0.34	0.10
p_{11}: CAD-Covered Area	0.75	0.75	0.25	0.25	0.75	0.75	0.25	0.75	0.56	0.26
p_{12}: CAD-Covered Area Alpha	1.00	0.94	0.10	0.10	0.21	0.78	0.10	0.38	0.45	0.39
p_{13}: CAD-Covered Area Size	0.66	0.36	0.20	0.20	0.26	0.32	0.20	0.42	0.33	0.16
p_{14}: Occluded Cars	0.75	0.75	0.25	0.75	0.75	0.75	0.75	0.25	0.63	0.23
p_{15}: Vehicle Status HUD	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.00
p_{16}: Vehicle Status HUD Alpha	0.10	0.13	0.54	0.10	0.20	0.72	0.36	0.32	0.31	0.23



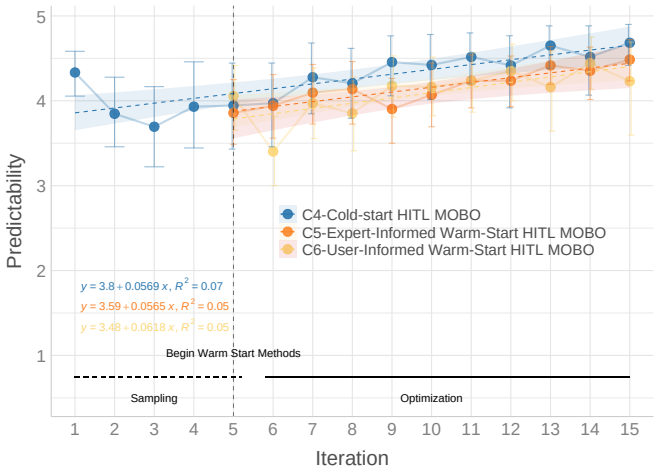
(a) Progression of perceived safety over the MOBO iterations.

(b) Progression of cognitive load over the MOBO iterations.

Figure 12: Value progression of perceived safety and cognitive load. The Warm-Start conditions had no sampling phase (i.e., iteration five was their first iteration) as they were initialized by the averaged expert design (in C5) or a custom user design (in C6).

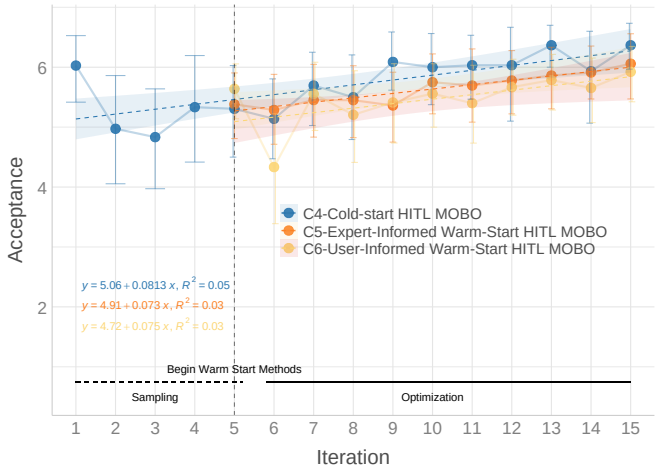


(a) Progression of trust values over the MOBO iterations.

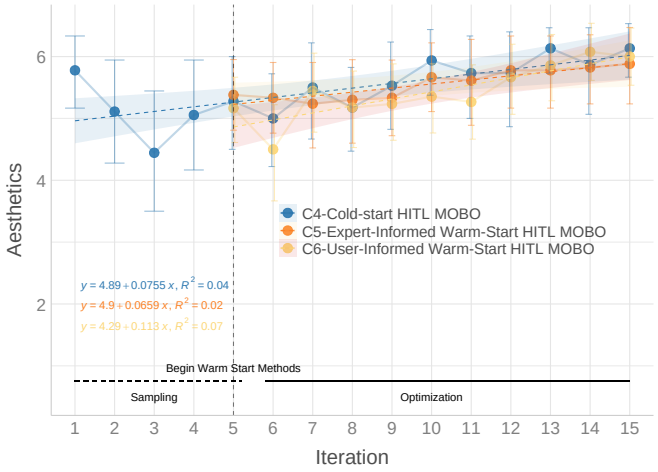


(b) Progression of predictability values over the MOBO iterations.

Figure 13: Value progression of trust and predictability.



(a) Progression of acceptance values over the MOBO iterations.



(b) Progression of aesthetics values over the MOBO iterations.

Figure 14: Value progression of acceptance and aesthetics.

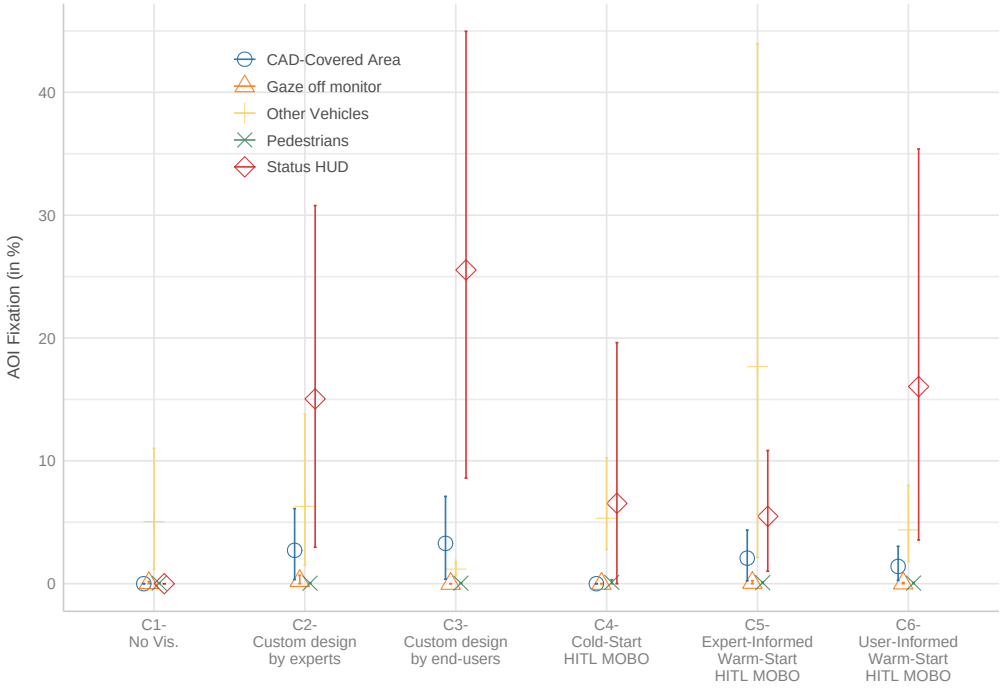


Figure 15: Eye fixations without the instances that the participants looked at the monitor but not at an AOI. For C6-User-Informed Warm-Start HITL MOBO, C2-Custom design by experts, and C3-Custom design by end-users, particular emphasis was placed on the speedometer. In C5-Expert-Informed Warm-Start HITL MOBO, emphasis was put on the car, which was also gazed upon comparatively frequently in the other conditions.