Teaching Multimodal Interaction in Cars to First-time Users

THOMAS MARINISSEN, Eindhoven University of Technology, The Netherlands
JONAS GLIMMANN, Mercedes-Benz AG, Germany
PAVLO BAZILINSKYY*, Eindhoven University of Technology, The Netherlands

This study explores three variations of a proactive method to teach multimodal gaze and gesture interactions to first-time users in the scenario of an SAE level 5 automated vehicle. The three variations differed in size, placement on the screen, and whether active user input was required to receive additional information. The results of a user study involving the gesture control prototype in a driving simulator (N = 30) show that the greatest variation was more effective in teaching, caused by significant differences in visibility ratings ($p < 0.001$), size ($p < 0.001$) and duration ($p = 0.001$) of the pop-ups. The results show no correlation between the measured effectiveness and the preference for a specific variation. Across all variations, participants are positive toward receiving proactive teaching from their car to learn new features. We conclude that proactively teaching users novel interaction methods has the potential to improve the user experience in future vehicles.

CCS Concepts: • **Human-centered computing** → **Gestural input**; **Empirical studies in interaction design**; **Accessibility technologies**.

Additional Key Words and Phrases: Automotive, User Interfaces, Teaching New Users, Multimodal Interaction, Gaze Control, Gesture Control

ACM Reference Format:
Thomas Marinissen, Jonas Glimmann, and Pavlo Bazilinskyy. 2018. Teaching Multimodal Interaction in Cars to First-time Users. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 16 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The car is transforming from a simple means of transportation into an automated driving (AD) and interconnected hub of technology. Firstly, there are notable developments in AD technology. The driver is transforming into a passenger, free to participate in various non-driving related tasks (NDRTs) such as relaxing or working [40, 43, 48]. The absence of a requirement for the driver to be in proximity to traditional driving controls (steering wheel, throttle, brake pedals, and gear selector) opens possibilities for diverse seating positions and configurations. For example, the driver could recline for a nap or swivel their seat 180°, creating a lounge-like experience with the rear passengers [36, 57]. This study assumes the context of cars reaching SAE level 5 AD in the future, where automated vehicles (AVs) operate without any driver intervention [19, 48, 50].

Authors' Contact Information: Thomas Marinissen, thomas.j.marinissen@gmail.com, Eindhoven University of Technology, Eindhoven, The Netherlands; Jonas Glimmann, jonas_david.glimmann@mercedes-benz.com, Mercedes-Benz AG, Sindelfingen, Germany; Pavlo Bazilinskyy, p.bazilinskyy@tue.nl, Eindhoven University of Technology, Eindhoven, The Netherlands.

Manuscript submitted to ACM

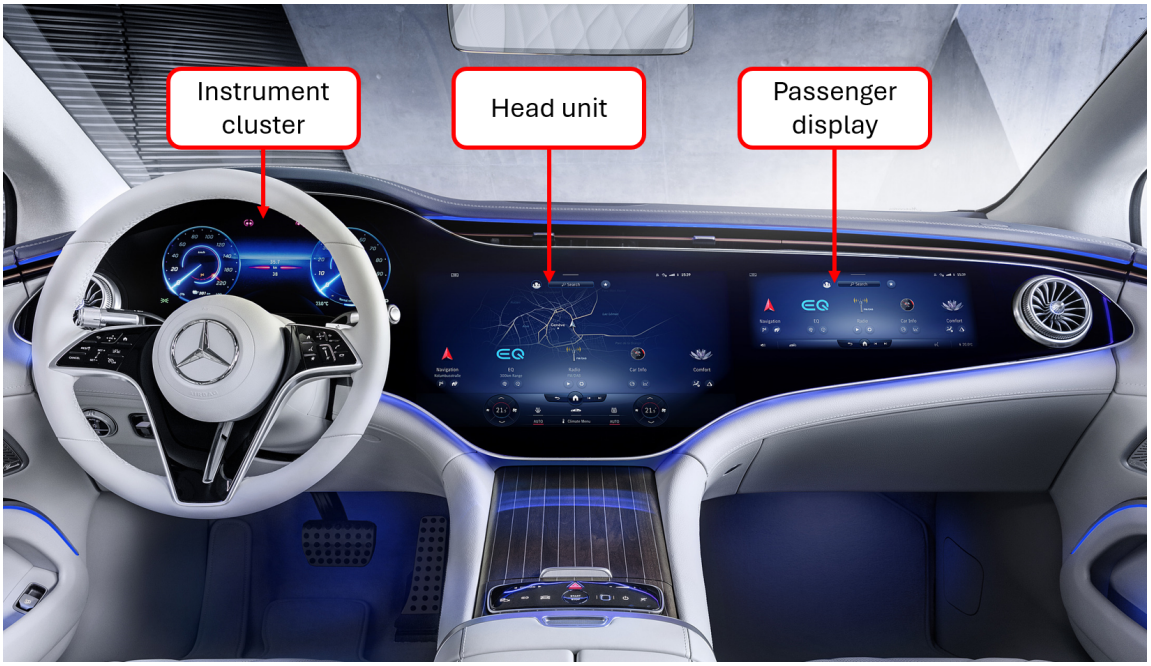

Fig. 1. Interior of the Mercedes-Benz EQS fitted with a three-screen layout [16].

Secondly, modern automotive user interfaces (UIs, see Figure 1 for an example) extend beyond traditional physical buttons and touchscreens by incorporating input modalities such as natural-language voice recognition [17, 18, 33, 37], eye gaze tracking [39], and gesture controls [12]. The risk of having multiple input modalities is that the user is overwhelmed by or even completely unaware of all the different options to operate the UI [29]. Creating awareness of and teaching novel interaction to new users could improve the car's user experience (UX).

1.1 Multimodal Interaction

Bourguet (2003) defines multimodal interaction as engaging with the virtual and physical environment through natural communication modes such as speech, gestures, or gaze [14]. Pfleging et al. (2012) explored speech and directional gesture controls on a touchpad for multimodal interaction, revealing high participant agreement on gesture commands [44]. Kern et al. (2010) investigated gaze tracking combined with a hard key on the steering wheel, which allows faster operation than speech recognition, but still slower than touchscreen use [26]. Gaze control is a potential safety concern in manual driving and is therefore more suitable in an SAE level 5 AV [48]. Both studies indicate potential clarity and consistency with novel interaction methods, focusing on quick user familiarization. However, they share a limitation: reliance on physical control elements within reach, unsuitable for varied seating positions in SAE level 5 AVs [48]. Automotive UI research explores diverse alternatives to traditional haptic controls, driven by the need for reachability and to accommodate the increasing functions in cars without a proportional rise in hard keys. While various combinations of modalities are studied, there isn't a universally superior choice yet.

Speech, gaze, and hand gestures, collectively known as natural interaction modalities, do not require reaching for control. Aftab (2019) suggests combining these to determine user intent [1]. Another study by Aftab & Von der Beeck

(2022) successfully identified various areas of the interior of a car, such as the steering wheel or gear selector, using gaze and gesture detection [2]. However, this set-up did not explore gaze detection in different areas of the screen or gestures beyond pointing. Many studies on multimodal interaction in vehicles focus on specific combinations of modalities without addressing how to teach this behavior to new users.

Natural interaction is already present in some cars and consumer electronics. For example, BMW introduced gesture controls in select models in 2015 [13]. In 2022, Li Auto released the L9 SUV, allowing passengers in the rear seat to control the entertainment display with advanced gesture recognition [49]. Audi's Urbansphere concept car, unveiled in the same year, showcases a multimodal interaction combining gaze and gesture control [9]. This concept inspired the prototype setup, mirroring the approach in which the eye gaze selects an item and gestures control or interact with the chosen feature. Similar multimodal interaction is observed in augmented reality (AR) glasses such as the Apple Vision Pro [6] and Meta Quest 3 [38]. Although AR glasses are currently niche products, their potential popularity could make this specific gaze and gesture interaction common social knowledge. However, before this happens, users need awareness and learning opportunities for these interactions.

1.2 Teaching and Learning of New Functions in Cars

Traditionally, learning about new functions in cars was done using a conventional printed user manual and/or by an explanation given by the salesperson in the dealership. Today, printed manuals are perceived as old-fashioned, difficult to navigate, frustrating, inefficient, and not sufficiently detailed [3, 41]. The explanation in the dealership must remain concise due to time constraints, which means that not all features of the cars are covered [10]. Users prefer to find the answers to their questions about their car online, asking people they know, in the digital version of the manual, or may even choose to keep problems unsolved [3, 41].

A common approach to modernizing the conventional user manual involves integrating it directly into the UI [20, 35, 51]. Similar indexing and explanations can be provided as in a regular user manual, with the added benefit of using interactive video and audio components [32, 34]. Another solution, as proposed by Alvarez et al. (2010), is voice-interfaced user help [3]. Virtual assistant technology, such as Apple Siri or Amazon Alexa, is already integrated in many modern cars from manufacturers such as Audi, BMW, Ford, General Motors, and Mercedes-Benz [4, 5, 8, 31, 33]. The benefit of this approach is that voice commands do not require the driver to take their eyes off the road and are therefore safe to use for users while driving the car. Studies show that users prefer interactive teaching methods to traditional user manuals [30, 41]. However, these interactive approaches to the user manual have one flaw in common with the traditional user manual: they require the user to initiate action to learn about new features. This means that features that the user is unaware of can remain undiscovered.

After interactive teaching, the next step to helping users find and learn new features is proactive teaching. Proactive teaching means that the car informs and instructs the user about new features by tracking which features have not yet been found and have not yet been used. This could be through audio messages or visual information pop-ups. The risk of this approach is that the user is interrupted in their activity, which can make teaching suggestions a nuisance rather than an aid. This means that it is important to consider the user's willingness to learn. This can be achieved by gently providing the right type and amount of information at an appropriate time. This is called *"nudging"* the user [15, 52, 54, 58]. This technique can be used to guide people towards a desired behavior, such as promoting healthy food [28]. Few automakers already use some form of nudging to attract user attention [12, 35]. Figures 2(a) and 2(b) show one example of an effective but subtle proactive approach to the teaching of gesture interactions by BMW. BMW cars equipped with this feature show information pop-ups, teaching the correct gesture controls [12, 13].

(a) Volume. (b) Answering calls.

Fig. 2. Information pop-up indicating gesture controls are available for different functions in the 2015 BMW 7-series [11, 12].

In Mercedes-Benz C-class and S-class, nudging is employed through on-screen information pop-ups, accessible as app icons, which explain new functions [32, 34, 35]. Users can create a personal account on various car brands to track undiscovered or unused functions. However, a drawback of nudging is that if the pop-ups are inconspicuous, users may not perceive or click them, leading to missed teaching information in the menu.

1.3 Aim of Study

This study examines the effectiveness and satisfaction of proactively teaching multimodal interaction to first-time users in cars. In the context of SAE level 5 AVs, with the possibility of varied seating configurations and with many car manufacturers equipping their vehicles with large or multiple screens, the reachability of controls could prove to be an issue. Examples of cars with multiple display areas are the Porsche Taycan, Honda E, and Mercedes-Benz EQS [16, 46, 55]. The assessed multimodal interaction, consisting of gaze and gesture input, addresses the challenges when touchscreens or physical controls are not within reach of the user. Users must be aware of and learn this novel interaction method. A user study was conducted to compare three variations of a proactive visual teaching method presented to users *"on the fly"* during regular interactions with the UI. We evaluated participants' awareness, interaction, adoption of presented information, and their preferences among the methods.

2 Method

2.1 Participants and Apparatus

30 residents of Germany participated in a user study between 8 August and 17 August 2023. All participants (14 female and 16 male) were older than 18 years old with a mean age of 45.7 years (SD = 14.7, median: 46.5) and had a driver's license. The participants, recruited through an external company that specializes in finding participants for user studies, were received in timeslots of two hours. The study was approved by the Ethics Review Board of Eindhoven University of Technology and the participants gave their informed consent to use their data.

The experiment was carried out on a seating buck, an immersive full-scale high-fidelity prototype of the interior of a luxury sedan to evaluate seating arrangements and new technologies during testing or development. Figure 3 shows a schematic drawing of the apparatus, in which the numbered elements represent the following components: (1) 3840x900

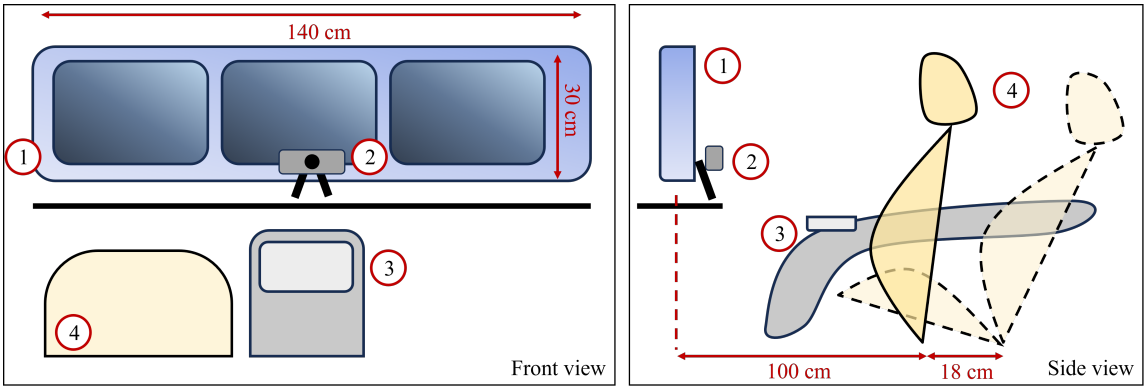

Fig. 3. Front and side view of the user study setup.

px widescreen, covering the entire width of the seating buck with a simple automotive GUI created specifically for this user study (as described in Section 2.2); (2) a centrally placed Intel RealSense [24] camera used for both gaze and gesture detection; (3) a center console with armrest and integrated Apple Magic Trackpad [7], used to remotely control the GUI as an alternative input method; (4) a remote-controlled electrically moving driver's seat. An upright and a laid-back seating positions were preprogrammed. For the laid-back seating position, the seat moves back approximately 18 cm and tilts backward to achieve a backrest angle of around 45°.

As an SAE level 5 AV scenario without driver intervention was assumed, driving controls such as the steering wheel and pedals were removed from the seating buck to enhance passengers' immersion in NDRTs. Gaze control, a potential safety concern in manual driving, is therefore more suitable for an SAE level 5 AV [48]. The camera (Item 2) captured eye gaze and gestures. Its video was fed to a computer, which generated a skeleton model of eyes, head, and right hand. This model was processed by a machine learning algorithm trained for this study to recognize gaze and gestures.

2.2 Concept

The concept in this study consists of two components: (1) a multimodal interaction formed by eye gaze and gesture detection to operate a prototype of an automotive graphical user interface (GUI) and (2) proactive teaching pop-ups that explain this multimodal interaction to first-time users.

The GUI used in the user study was created in Protopie [53]; Figure 5 for the initial version (without the information pop-up). It was designed specifically for widescreen. The screen layout is designed to be like modern car GUIs with three screens, consisting of (from left to right, as seen in Figure 1 [16]) an instrument cluster (IC), head unit (HU) and passenger display (PD) [16]. This means that each window in our GUI can be placed in one of these three positions. Two windows were made interactable in this prototype: a music player and a video player, which could play one pre-loaded song and video, respectively. Both windows supported all functions presented in Figure 5 and could be placed in the IC, HU, or PD area. The IC was directly in front of the participant (or driver, in real life) and allowed an optimal viewing position.

The combination of gaze and gesture in our prototype was inspired by AR glasses. Eye gaze was used to determine which window on the screen the user wanted to operate, and gestures were used to interact with (the functions of) the selected window. These specific hand gestures were chosen because the system recognized them reliably. The gaze

(a) **Audio control**: Thumbs down gesture for *volume down.*

(b) **Audio control**: Thumbs up gesture for *volume up.*

(c) **Audio control**: Index finger on lips gesture for *mute.*

(d) **Content timing control**: Pinch and drop gesture for f*ast forward/reverse.*

(e) **Content timing control**: Hand up gesture for *pause.*

(f) **Content timing control**: Wave twice gesture for *start of play.*

(g) **Content timing control**: Wave twice gesture for *end of play.*

(h) **Screen layout**: Grab gesture for *start of relocating window.*

(i) **Screen layout**: Drop gesture for *end to relocating window.*

Fig. 4. Frames of the animations for gesture categories: audio control (a—c), content timing control (d–g), and screen layout (h–i).

detection system of the prototype allowed us to determine whether users were looking at the IC, HU, PD, or off the screen. If the gaze was detected to be on a section of the screen with a window, the gesture input was used to control only that specific window. One goal of this was to reduce false positives. To inform the user about which window gaze was being detected, the system provided the user with visual feedback: a colored outline appeared around the window if it was looked at.

For gestures, the feedback depended on which gesture was performed. “Play”, “Pause”, and the three audio functions were shown with a pictogram on screen. “Forward” and “Reverse” were displayed by a time stamp indicator, similar to a slider, which doubled in size and turned orange when activated. This indicator moved according to whether the user moved their hand right (“Forward”) or left (“Reverse”). Feedback for “Relocate Windows” consisted of three steps. Firstly, the “grab” gesture triggered the selected window to *pop out* and enlarge. Secondly, the window would follow the user’s hand movement across the horizontal axis, allowing the user to determine where the window should be placed. Lastly, the window was *dropped* if the user opened their hand. It then shrunk back to its original size and slid *magnetically* into place on the IC, HU, or PD, depending on where the user left it closest to. Any window previously in that position was automatically moved to another position.

The proactive teaching of how to operate one of the functions through multimodal interaction consisted of (1) animations that showed a person performing the gestures correctly, (2) textual explanations, and (3) pictograms. The static frames of the animations can be found in Figure 4. An animation cycle lasted 3 s (for volume and mute) or 6 s (for play, pause, fast forward/reverse, and window relocation) and was looped until the entire pop-up disappeared from the screen. The three conditions varied in the following aspects: (1) how the teaching information was accessed, (2) the size of the pop-ups, (3) where the information pop-ups appeared on screen, and (4) how many animations could appear simultaneously on one slide of a pop-up. For all conditions, the animations were grouped according to the categories found in Figure 4, which means that each condition had one (set of) pop-up(s) for each category of functions.

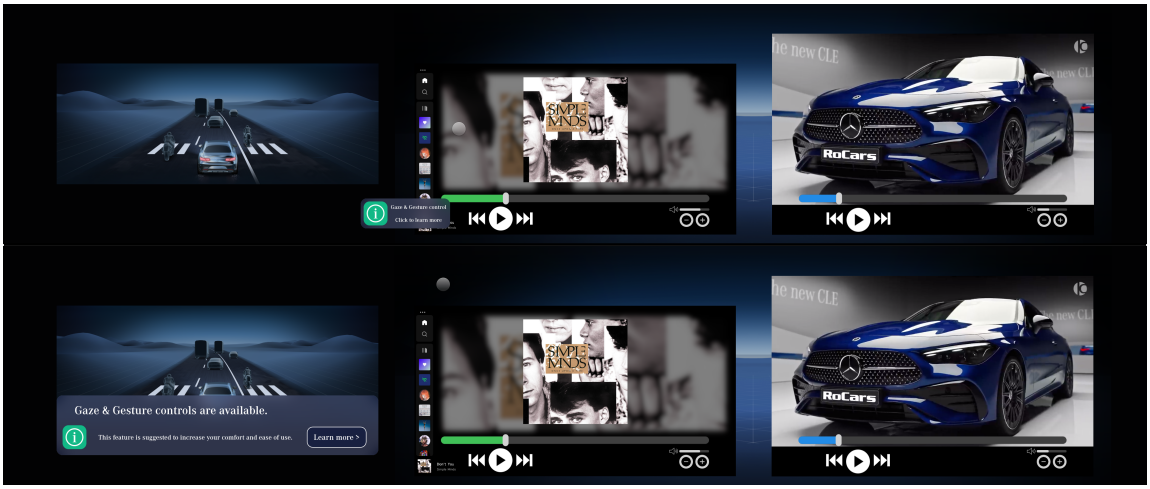

Fig. 5. Information pop-ups used in the study: C1 (top) shows a small pop-up located on the side of the HU, while C2 (bottom) uses a larger, centrally placed pop-up for improved visibility.

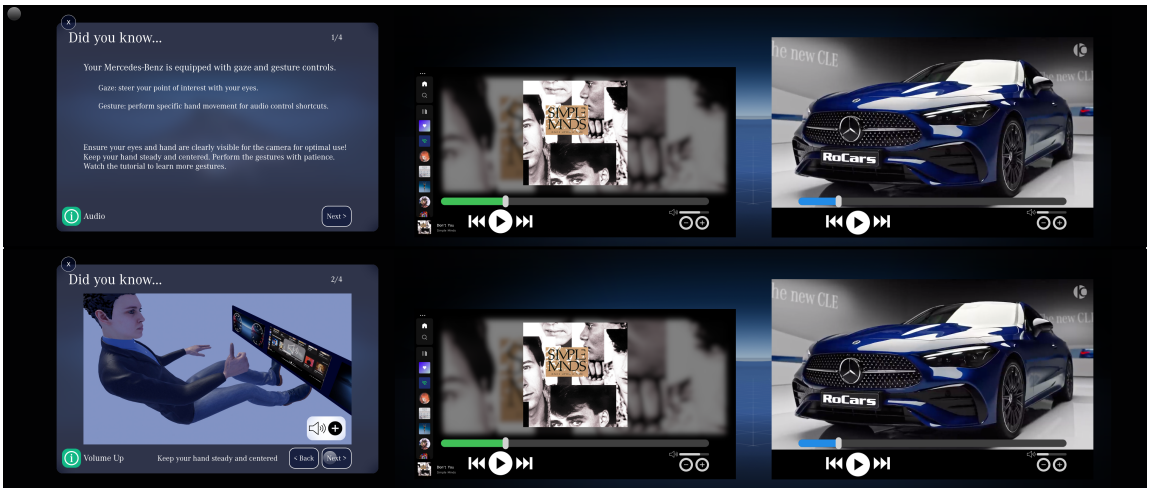

Fig. 6. Slides from the information menu used in C1 and C2: textual explanation (top) and animated feedback (bottom).

In the study, three conditions were tested, each differing in the size, placement, and interactivity of the instructional pop-ups. *Condition 1* (C1; tiny) featured the smallest pop-up (250x80 px) positioned on the side of the head unit (HU), as shown in Figure 5. It indicated the availability of gaze and gesture controls. Users could access additional instructional content, including text and animations, by clicking the pop-up via the touchpad (Item 3 in Figure 3). If left untouched, the pop-up disappeared after 30 s, and the teaching content remained inaccessible. This condition served as the baseline, resembling current implementations in Mercedes-Benz vehicles [32, 34]. *Condition 2* (C2; medium) operated similarly to C1 but used a larger pop-up (1200x200 px), centrally placed to increase visibility (Figure 5). Both C1 and C2 provided access to the same interactive information and animation slides, seen in Figure 6, which users could browse at their own pace. *Condition 3* (C3; big) offered the largest instructional content (ranging from 600x500 to 1500x500 px) and differed fundamentally in interaction. Pop-ups appeared automatically and directly on the instrument cluster (IC), requiring no user input to access. As shown in Figure 7, the animations began to play immediately after recline of the seat and disappeared after 30 s. Depending on the category, Timing, Audio, or Screen Layout, pop-ups could show one to three animations simultaneously. C3 was entirely passive, prioritizing visibility and immediacy over user-controlled exploration.

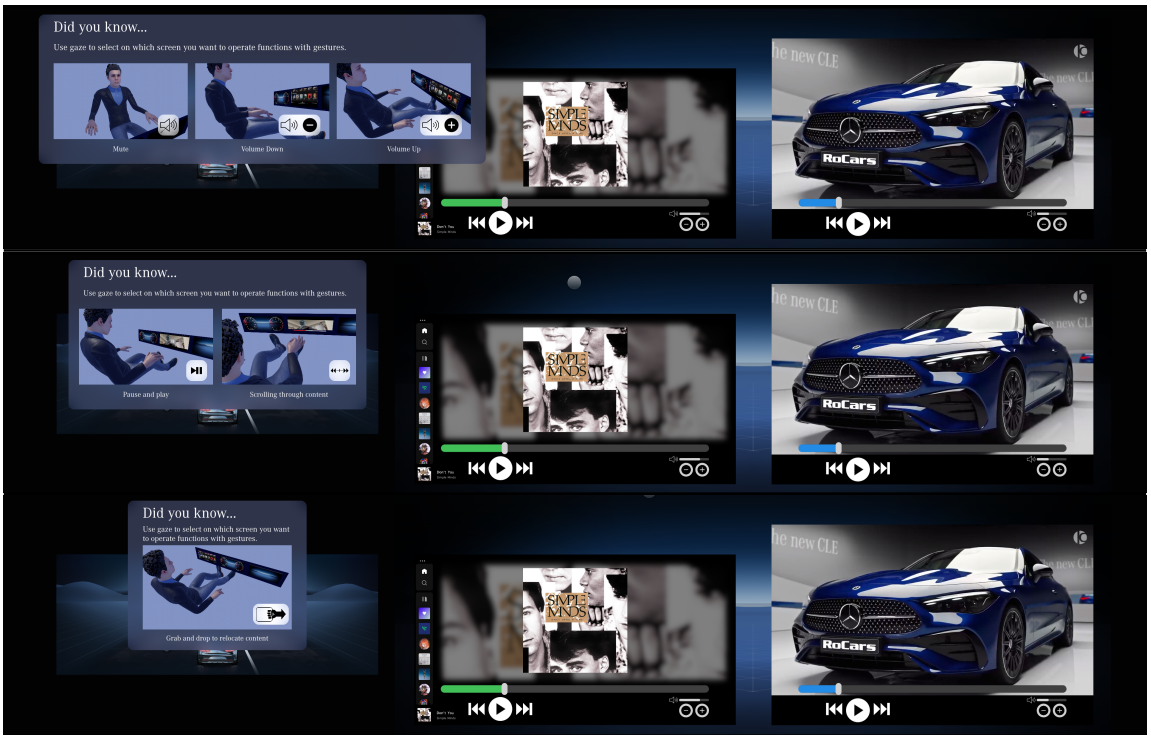

Fig. 7. Information pop-ups in C3 explaining gesture controls for different function categories: audio (top), content timing (middle), and screen layout (bottom).

2.3 Procedure and Data Analysis

The user study consisted of two parts designed to evaluate how users experience and learn multimodal interaction in the context of an SAE Level 5 automated vehicle (AV) [48]. The age and gender of the participants were recorded at the start of the study.

In Part 1, participants were gradually introduced to the concept of multimodal interaction without being explicitly informed about the objective of the study. To simulate a naturalistic non-driving-related task (NDRT) scenario, participants were seated in a reclined position and asked to engage in a structured secondary task. First, they watched a widescreen video that showcased the UI features of the Mercedes-Benz CLE (Item 1 in Figure 3) [47]. Next, they read a one-page text describing user interfaces in current Mercedes-Benz models. Finally, they re-watched the video (included in the supplementary material). The order of these tasks remained consistent for all participants. The examiner explained that the secondary task served two purposes: (1) to evaluate the comfort and realism of the reclined seat position and (2) to identify an incorrect statement in the text by comparing it with the video content, presented as a challenge to encourage engagement. While participants were immersed in this secondary task, they experienced three remotely activated seat recline movements. Each recline triggered a pop-up associated with a specific function category (see Figure 4), all under a single randomly assigned teaching condition. Pop-ups were displayed on the GUI for 30 s while participants continued the secondary task. Afterwards, they were prompted to use the functions in the shown category, without being told which input modality to use. The examiner then returned the seat to its upright position. This cycle was repeated for all three categories of functions. Ten participants were able to experience each condition, ensuring a total of 30 exposures per function category pop-up. Participant behavior during Part 1 was assessed through direct observation. The examiner recorded whether participants (1) noticed the pop-up, (2) interacted with it (e.g., attempted to click), and (3) perceived gaze feedback. If a participant did not perceive the pop-up content, learning and gesture adoption were considered unlikely.

In Part 2, participants were informed about the true objective of the study and the concept of multimodal interaction. They were given time to review the pop-ups from Part 1 and freely try the interaction techniques. They then completed a series of questionnaires that evaluated the teaching method they had just experienced. Finally, they reviewed the two remaining teaching conditions (which they had not encountered before), allowing comparative feedback. The session ended with a set of open questions to gather qualitative impressions.

After each condition, the participants were instructed to assess the teaching method, not the technical quality of the prototype. They responded to three scales: (1) a 20-point *NASA TLX scale* to measure perceived workload [21], (2) *KANO scale* to evaluate reactions to novel features, processed according to the original method [25] and (3) *Acceptance scale* for perceived usefulness and satisfaction on a 5-point scale (from -2 to +2) [56]. In addition, a questionnaire assessed the visual properties and content of the information pop-ups using a 7-point Likert scale.It included 12 items: Q1–Q3 addressed animations, clarity, and visibility; Q4–Q5 assessed pop-up size and duration; Q6–Q7 focused on informativeness and task interference; Q8–Q11 explored the timing of pop-ups, preference for animation versus text, and openness to proactive feature learning; Q12 provided general feedback on system guidance. Questions Q1–Q7 were answered after each condition, while Q8–Q12 were presented only after the first condition. At the end of the study, participants ranked the three teaching conditions from most to least preferred and selected their favorite and least favorite gesture. See the supplementary material for the forms used.

	C1	C2	C3
First appearance	5⟶ 0	8⟶ 0	10⟶ 4
Second appearance	1⟶ 0	3⟶ 1	8⟶ 4
Third appearance	4⟶ 0	8⟶ 0	10⟶ 4
Total	10⟶ 0	19⟶ 1	28⟶ 12
p_n = 0.72			

Table 1. Results ($n \longrightarrow i$) of participant's notices count (n) and interaction count (i) with each appearance of an information pop-up for all three conditions. p_n is the p-value result of the Chi-Square test for n.

For the questionnaire, NASA TLX scale, and Acceptance scale, an ANOVA test was used to determine significance. For the ranking of the conditions, a Chi-Pearson square test was used to determine significance. For all tests in this paper, an alpha level of 0.05 was used. Data analysis was performed in Microsoft Excel.

3 Results

3.1 Notices and Interaction Rates

The data of all participants was retained as there was no need to filter out the participants. Table 1 displays the notices and interactions for each condition. The results show that C3 was noticed the most often of the three conditions (N = 28) and C1 the least frequently (N = 10). A Chi-Pearson square test was performed under the three conditions, which did not produce significant differences in the number of notices. For all conditions, the second appearance notices, during which participants performed the reading task, resulted in lower notices (N = 1,3,8 for C1, C2 and C3, respectively) than for the first (N = 5,8,10 for C1, C2 and C3, respectively) and third (N = 4,8,10 for C1, C2, and C3, respectively). C1 did not produce interactions. For this condition, the user had to click on the information notification to reach the full explanation of gaze and gesture control. The complete lack of interactions for this condition implies that none of the participants in this group learned about the gaze and gesture control that was available. C2 produced only one interaction of the 30 possible interactions between the participants. C3 had 12 interactions out of 30 possible interactions, resulting in the highest interaction rate of the three conditions.

3.2 Questionnaires and Scales

Table 2 shows the results of the ANOVA test for the questionnaire. For dimensions Q3, Q4 and Q5 (visibility, size, and duration), the results show significant differences between the three conditions and are marked in bold. The other dimensions did not show any significant differences. Table 3 displays the results of the ANOVA tests for the NASA TLX and the Acceptance scale. The results show that there was a significant difference, marked in bold, only for the dimension of temporal demand. In this dimension, C2 scored the lowest demand. No significant differences were found in any of the other five dimensions of the NASA TLX. Table 3 also shows that there were no significant differences in usability or satisfaction dimensions for the Acceptance scale. Table 4 shows the results of the Kano scale.

3.3 Ranking of Conditions and Gestures

Table 5 shows how participants ranked the three conditions. C3 and C2 were the most popular teaching methods, and C1 was the least favorite. The p-value indicates that these results differed significantly from the expected values. The participants indicated their favorite and least favorite of the six different types of gestures, of which the results are

Question	M			p
	C1	C2	C3	
Q1 animations	5.63	5.87	5.57	0.60
Q2 understandable	5.37	6.07	5.97	0.13
Q3 visibility	4.47	5.97	6.13	**<0.001**
Q4 size	2.53	4.60	4.17	**<0.001**
Q5 duration	3.60	4.33	4.17	**0.001**
Q6 enough info	5.73	6.03	6.03	0.55
Q7 interference	3.77	4.43	4.67	0.11
Q8 moment	3.40	4.80	4.20	0.19
Q9 windows	6.40	5.50	6.50	0.22
Q10 descriptions	2.70	3.20	1.90	0.19
Q11 proactive	5.80	6.30	6.20	0.46
Q12 feedback	5.60	5.20	5.90	0.55

Table 2. Results of the ANOVA test on all three conditions across the twelve dimensions of the questionnaire (on a 7-point Likert scale (from 1–7)).

	Mean value			p
NASA TLX	**C1**	**C2**	**C3**	
Mental demand	8.03	6.97	7.93	0.42
Physical demand	3.70	4.30	3.87	0.32
Temporal demand	7.73	5.37	7.23	**0.02**
Performance	6.27	5.50	6.33	0.35
Effort	7.10	6.70	7.63	0.52
Frustration	4.80	5.00	5.53	0.63
Acceptance scale				
Usefulness	1.20	1.29	1.28	0.48
Satisfying	1.16	1.26	1.26	0.62

Table 3. Results of the ANOVA test on the conditions across the six dimensions of the NASA TLX scale (on a 20-point scale 1–20) and usefulness and satisfaction of the Acceptance scale (on a 5-point scale -2–+2)) [21, 56].

Feature	C1	C2	C3
Performance	8	11	10
Must-have	13	13	11
Attractive	1	0	2
Indifferent	7	4	5
Reverse	1	1	1
Questionable	0	1	1

Table 4. Results of the Kano scale for all three conditions [25].

shown in Table 6. The results show that the mute gesture was voted as the favorite (N = 14). Pinch and play were rated as the least favorite gestures (N = 11). Neither pinch nor play received any votes for being a favorite gesture.

Ranking	C1	C2	C3
1^{st} (favorite)	5	12	13
2^{nd}	9	10	11
3^{rd} (least favorite)	16	8	6
p = **0.048**			

Table 5. Results for the ranking of the three conditions from favorite to least favorite.

	Mute	Volume	Grab	Pause	Pinch	Play
Favorite	14	7	6	2	0	0
Least fav.	1	3	3	0	11	11

Table 6. Results for rating of gestures.

4 Discussion

In this paper, we examined a proactive approach to teaching first-time users of gaze and gesture control in a driving scenario in an SAE level 5 AV. The study, conducted among 30 participants performing a secondary task (reading or watching video) in a dedicated seating buck, presented animations and textual information while the participants reclined. Three variations (C1–C3) of this teaching method were tested, varying in size, placement on the screen, and interaction requirements. The aim was to explore participants' preferences and effectiveness in receiving proactive instructions and to find the most preferred and effective of the three variations.

The results of the user study reveal that C3 proved to be more effective in notifying participants about gaze and gesture technology, leading to the highest number of learned interactions. C2 resulted in only one interaction, while C1 had none. The discrepancy in notices and interactions can be attributed to the greater prominence of C3 due to its larger size and more contrasting colors, in agreement with previous research on visual significance and attention in human-computer interaction [23, 59], suggesting that the larger size and central placement of C3 enhanced noticeability. The automatic presentation of C3 without requiring user interaction further contributed to its effectiveness, consistent with the literature on passive guidance and low effort engagement [42]. The importance of pop-up size (Q3) and visibility (Q4) importance is reinforced by the significant differences in ratings for each condition ($p_{Q3} < 0.001$ and $p_{Q4} < 0.001$, respectively). Surprisingly, while the large pop-ups of C3 and C2 covered a substantial section of the IC, there were no significant differences in how they interfered with the secondary task, as indicated by Q7. Strangely, the results of Q5 (duration) and the temporal demand of NASA TLX indicate significant differences ($p_{Q5} = 0.001$ and $p = 0.02$, respectively) in the duration ratings of pop-up appearance, despite being fixed to 30 s. This was probably caused by the pressure participants felt to read the text or learn the gestures quickly.

Although C3 is the most effective in teaching multimodal interaction, the results of the acceptance scale do not suggest significant differences in perceived usefulness or satisfaction among the participants. This discrepancy likely stems from the participant rating the *concept* of proactive teaching rather than its specific implementation. However, positive ratings are evident for both dimensions under all conditions. This aligns with findings in the trust and automation acceptance literature [22, 27]. The Kano scale and Q9–Q12 further affirm the positive evaluation of the proactive teaching concept by the participants, with the 'must have' category consistently scoring highest and the 'reverse' category scoring very low for each condition.

C1 was significantly less favored ($p = 0.048$) than C3 and C2, while C3 and C2 are consistently ranked as favorites over C1. This outcome can be attributed to several factors. Firstly, C3 and C2 garnered the most attention, which was perceived positively. In contrast, C1's lower visibility due to its small size contributed to its least favorite status. Secondly, participants appreciated C3's immediate animations and C2's clear indication of a clickable menu, advantages lacking in C1. Furthermore, participants valued the ability to close the info pop-up in C1 and C2. The results underscore the participants' inclination for clear, proactive teaching features in their cars.

Table 6 reveals that the Mute gesture is the most popular, likely due to its ease of execution, reliable recognition, and widespread social familiarity. The participants quickly grasped the connection between the hand movement and its controlled function, aided by the straightforward animation. In contrast, the Play and Pinch gestures were least favored, as precise execution posed difficulties for the system. Participants struggled with recognition and did not associate hand movements with their respective functions. Complex animations, particularly for the Play and Pause gestures, led to confusion as both functions were combined in a single video. In summary, simple, focused animations aligned with the gesture's function proved effective, highlighting the importance of clear associations for user understanding.

4.1 Limitations and Future Work

This study is limited by the short time participants had to practice and assimilate multimodal interaction. In real-world scenarios, users would engage with such systems for extended periods, allowing more natural learning curves. Future research should explore longitudinal effects and iterative teaching strategies. Moreover, while our prototype focused solely on gaze and gesture input, integrating voice control could expand usability, particularly in scenarios where touch is impractical [45]. System limitations include occasional gesture recognition errors and low gaze resolution: only broad focus zones (IC, HU, PD, or off-screen) were distinguishable. More accurate gaze tracking, such as that found in the Apple Vision Pro [6], could improve precision and reduce false positives. The prototype also lacked automatic logging of user interactions, gaze data, or system timestamps, which would have allowed more robust data analysis. Lastly, questionnaire analysis was performed using Microsoft Excel. Employing more advanced statistical tools could increase analytical accuracy. We encourage future studies to explore richer input modalities, improved data tracking, and evaluate proactive teaching strategies over longer time frames in real-life settings.

5 Conclusion

The main conclusion of this study is that the participants are positive about receiving proactive instruction from their car, regardless of the variation in the information pop-up shown. The majority of participants think proactive teaching is either a must-have or a performance feature in a car that can be controlled by new (multimodal) interactions. C3 was measured to be the most effective teaching method, C1 and C2 were not effective. The discrepancy in effectiveness is caused by significant differences in the visibility, size, and duration ratings of the pop-ups. Future cars should provide users with the option to learn new features with clear and visible instructions.

6 Supplementary Material

Supplementary material containing materials used in the user study is available at: https://dropbox.com/scl/fo/phkpuxl12eeils4ylxlvx/AOZ-2nRVXgCOuVf3rugx0_A?rlkey=8oo0oc38hleyic8mxnabcl5ko.

References

[1] Abdul Rafey Aftab. 2019. Multimodal driver interaction with gesture, gaze and speech. In *Proceedings of the 2019 International Conference on Multimodal Interaction (ICMI '19)*. ACM, Suzhou, China, 487–492. https://doi.org/10.1145/3340555.3353726

[2] Abdul Rafey Aftab and Michael von der Beeck. 2022. Multimodal driver referencing: A comparison of pointing to object inside and outside the vehicle. In *Proceedings of the 27th International Conference on Intelligent User Interfaces (IUI '22)*. ACM, Helsinki, Finland, 483–495. https://doi.org/10.1145/3490099.3511122

[3] Ignacio Alvarez, Aqueasha Martin, Jerone Dunbar, Joachim Taiber, Dale-Marie Wilson, and Juan E. Gilbert. 2010. Voice interfaced user help. In *Proceedings of the Second International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI 2010)*. ACM, Pittsburgh, PA, USA, 42–49. https://doi.org/10.1145/1969773.1969783

[4] Amazon. 2024. Vehicles with Alexa. https://www.amazon.com/alexa-auto/b?ie=UTF8&node=17744356011. Accessed: 2024-01-20.

[5] Amazon Developers. 2024. What is Alexa? https://developer.amazon.com/en-GB/alexa#:~:text=Alexa%20is%20Amazon's%20cloud%2Dbased,technology%20they%20use%20every%20day. Accessed: 2024-01-20.

[6] Apple. 2024. Apple Vision Pro. https://www.apple.com/apple-vision-pro. Accessed: 2025-04-08.

[7] Apple. 2024. Magic trackpad–white multi-touch surface. https://www.apple.com/shop/product/MK2D3AM/A/magic-trackpad-white-multi-touch-surface. Accessed: 2024-03-12.

[8] Apple. 2024. Siri. https://www.apple.com/siri. Accessed: 2024-01-20.

[9] Audi. 2022. Space travel in the heart of the megacity. https://www.audi-mediacenter.com/en/press-releases/space-travel-in-the-heart-of-the-megacity-14595. Accessed: 2024-01-03.

[10] Emilio Bellini, Claudio Dell'Era, Federico Frattini, and Roberto Verganti. 2017. Design-driven innovation in retailing: An empirical examination of new services in car dealership. *Creativity and Innovation Management* 26 (2017), 91–107. Issue 1. https://doi.org/10.1111/caim.12140

[11] BimmerTech. 2020. BMW gesture control–the next level of idrive interaction. https://www.bimmer-tech.net/blog/item/124-bmw-gesture-control. Accessed: 2024-01-20.

[12] BMW Group. 2015. The new BMW 7 Series. https://www.press.bmwgroup.com/global/article/detail/T0221224EN/the-new-bmw-7-series. Accessed: 2024-01-20.

[13] BMW Group. 2019. Get the most out of gesture control–BMW how-to. https://www.youtube.com/watch?v=_mGwJh4da5w. Accessed: 2024-01-03.

[14] Marie-Luce Bourguet. 2003. Designing and prototyping multimodal commands. In *Human-Computer Interaction–INTERACT'03*. IOS Press, Zurich, Switzerland, 717–720.

[15] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow, Scotland, UK, 1–15.

[16] Caricos. 2022. 2022 Mercedes-Benz EQS–interior, cockpit. https://www.caricos.com/cars/m/mercedes-benz/2022_mercedes-benz_eqs/images/58.html. Accessed: 2024-01-23.

[17] Cerence. 2021. How Mercedes is creating innovative multi-modal experiences with cerence look. https://www.cerence.com/news-releases/news-release-details/how-mercedes-creating-innovative-multi-modal-experiences-cerence/. Accessed: 2024-01-20.

[18] Cerence. 2023. Gaze detection. https://www.cerence.com/cerence-products/apps-multi-modality. Accessed: 2024-01-20.

[19] Ching-Yao Chan. 2017. Advancements, prospects and impacts of automated driving systems. *International Journal of Transportation Science and Technology* 6, 3 (2017), 208–216. https://doi.org/10.1016/j.ijtst.2017.07.008

[20] Google Cloud. 2021. From print to voice and beyond: how toyota is transforming the car manual. https://cloud.google.com/blog/topics/manufacturing/toyota-modernizes-the-car-manual-with-google-cloud. Accessed: 2024-01-20.

[21] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human Mental Workload, Advances in Psychology* 52 (1988), 139–183. https://doi.org/10.1016/B978-0-08-037602-5.50034-3

[22] Kevin A Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors* 57, 3 (2015), 407–434.

[23] William J Horrey and Christopher D Wickens. 2006. Driver workload and performance effects of adaptive cruise control and crash warning systems. *Human Factors* 48, 3 (2006), 682–692.

[24] Intel. 2024. Intel Realsense. https://www.intelrealsense.com/. Accessed: 2024-03-12.

[25] Noriaki Kano, Nobuhiku Seraku, Fumio Takahashi, and Shinichi Tsuji. 1984. Attractive quality and must-be quality. *Journal of the Japanese Society for Quality Control* 14, 2 (1984), 39–48. https://doi.org/10.1007/BF03189671

[26] Dagmar Kern, Angela Mahr, Sandro Castronovo, Albrecht Schmidt, and Christian Mueller. 2010. Making use of drivers' glances onto the screen for explicit gaze-based interaction. In *Proceedings of the Second International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI 2010)*. ACM, Pittsburgh, PA, USA, 110–113. https://doi.org/10.1145/1969773.1969796

[27] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 1 (2004), 50–80.

[28] Min Kyung Lee, Sara Kiesler, and Jodi Forlizzi. 2011. Mining behavioral economics to design persuasive technology for healthy choices. In *CHI '11: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Vancouver, BC, Canada, 325–334.

[29] Lentz, Alison and Schlesinger, Benny and DiMartile III, John Thomas and Taubman, Gabriel and O'Dell Regina. 2018. A logical layer to interpret user interactions. https://www.tdcommons.org/dpubs_series/1223/. Accessed: 2024-01-23.

[30] Tomas Macek, Martin Labsky, Jan Vystrcil, David Luksch, Tereza Kasparova, Ladislav Kunc, and Jan Kleindienst. 2014. Interactive car owner's manual user study. In *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '14)*. ACM, Seattle, WA, USA, 1–4.

[31] Mercedes-Benz Group AG. 2018. It understands you perfectly. https://group.mercedes-benz.com/company/magazine/technology-innovation/mbux-voice-assistant-hey-mercedes.html. Accessed: 2024-01-20.

[32] Mercedes-Benz Group AG. 2023. C-class saloon manual–calling up the digital owner's manual. https://www.mercedes-benz.co.uk/passengercars/services/manuals.html/c-class-saloon-2023-09-w206-mbux/digital-owners-manual/calling-up-the-digital-owners-manual. Accessed: 2024-01-22.

[33] Mercedes-Benz Group AG. 2023. Mercedes-Benz takes in-car voice control to a new level with ChatGPT. https://group.mercedes-benz.com/innovation/digitalisation/connectivity/car-voice-control-with-chatgpt.html. Accessed: 2024-01-20.

[34] Mercedes-Benz Group AG. 2023. S-class saloon manual–calling up the digital owner's manual. https://www.mercedes-benz.co.uk/passengercars/services/manuals.html/s-class-saloon-2023-09-w223-mbux/digital-owners-manual/calling-up-the-digital-owners-manual. Accessed: 2024-01-22.

[35] Mercedes-Benz Group AG. 2024. Discover your owner's manual. https://www.mercedes-benz.co.uk/passengercars/services/manuals.html. Accessed: 2024-01-22.

[36] Mercedes-Benz USA. 2015. The Mercedes-Benz F 015 luxury in motion. https://media.mbusa.com/releases/the-mercedes-benz-f-015-luxury-in-motion. Accessed: 08-04-2025.

[37] Mercedes-Benz USA. 2020. Meet the S-class digital: My mbux (Mercedes-Benz user experience). https://media.mbusa.com/releases/release-9e110a76b364c518148b9c1ade19bc23-meet-the-s-class-digital-my-mbux-mercedes-benz-user-experience. Accessed: 2024-01-20.

[38] Meta Quest. 2023. This is Meta Quest 3. https://www.youtube.com/watch?v=Exu7r2vZpcw. Accessed: 2024-01-03.

[39] MotorTrend. 2023. The 2024 bmw 5 series lets you steer with just your eyes. https://www.motortrend.com/news/2024-bmw-5-series-eye-lane-change-tech/. Accessed: 2024-01-20.

[40] Frederik Naujoks, Dennis Befelein, Katharina Wiedemann, and Alexandra Neukum. 2017. A review of non-driving-related tasks used in studies on automated driving. In *Advances in Human Aspects of Transportation*. Springer, Los Angeles, CA, USA, 525–537.

[41] David G. Novick and Karen Ward. 2006. Why don't people read the manual. In *SIGDOC '06: Proceedings of the 24th annual ACM international conference on Design of communication*. ACM, Myrtle Beach, SC, USA, 11–18.

[42] Raja Parasuraman and Mustapha Mouloua. 2000. Adaptive automation and human performance: A literature review. *Human Factors* 42, 2 (2000), 379–401.

[43] Bastian Pfleging, Andrew L. Kun, and Orit Shaer. 2022. Automated vehicles as a space for work & wellbeing. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*. Association for Computing Machinery, New Orleans, LA, USA. https://doi.org/10.1145/3491101.3503766

[44] Bastian Pfleging, Stefan Schneegass, and Albrecht Schmidt. 2012. Multimodal interaction in the car: Combining speech and gestures on the steering wheel. In *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '12)*. ACM, Portsmouth, NH, USA, 155–162. https://doi.org/10.1145/2390256.2390284

[45] Cecily Pickering, Gary Burnett, and Julie A. Hansen. 2021. Voice user interfaces in vehicles: A review of the current state and future directions. *International Journal of Human-Computer Studies* 148 (2021), 102580.

[46] Porsche AG. 2024. Porsche Taycan. https://www.porsche.com/international/models/taycan/taycan-models/taycan/. Accessed: 2024-03-08.

[47] RoCars (Youtube channel). 2023. 2024 Mercedes Cle–Nww coupe from Mercedes in details. https://www.youtube.com/watch?v=ag9NQMrXM00. Accessed: 2024-01-03.

[48] SAE International. 2021. SAE levels of driving automation refined for clarity and international audience. https://www.sae.org/blog/sae-j3016-update. Accessed: 2024-01-20.

[49] Screens. 2023. Li auto L9 gesture control. https://www.youtube.com/watch?v=q09VOemNji8. Accessed: 2024-01-03.

[50] Kong Joo Shin, Naoto Tada, and Shunsuke Managi. 2019. Consumer demand for fully automated driving technology. *Economic Analysis and Policy* 61 (2019), 16–28. https://doi.org/10.1016/j.eap.2019.01.002

[51] Skoda Auto A.S. 2023. Digital manual. https://www.skoda-auto.com/connectivity/infotainment-apps-digital-manual. Accessed: 2024-01-20.

[52] Carola Stryja and Gerhard Satzger. 2018. Digital nudging to overcome cognitive resistance in innovation adoption decisions. *The Service Industries Journal* 39 (2018), 1123–1139. Issue 15-16. https://doi.org/10.1080/02642069.2018.1485767

[53] Studio XID. 2024. Protopie website. https://www.protopie.io/. Accessed: 08-04-2025.

[54] Cass Sunstein and Richard Thaler. 2008. Nudge: Improving decisions about health, wealth and happiness. https://www.researchgate.net/profile/C-Sunstein/publication/235413094_NUDGE_Improving_Decisions_About_Health_Wealth_and_Happiness/links/00b49534135e982d0a000000/NUDGE-Improving-Decisions-About-Health-Wealth-and-Happiness.pdf. Accessed: 2024-01-23.

[55] Top Gear Magazine. 2020. Honda E review. https://www.topgear.com/car-reviews/honda/e. Accessed: 2024-03-08.

[56] 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. https://doi.org/10.1016/S0968-090X(97)00002-9

[57] Volvo Cars. 2015. Volvo cars unveils concept 26, delivering the luxury of time. https://www.media.volvocars.com/global/en-gb/media/pressreleases/169396/volvo-cars-unveils-concept-26-delivering-the-luxury-of-time. Accessed: 08-04-2025.

[58] G. Wessel, E. Alterndorf, M. Schwalm, Y. Canpolat, C. Burghardt, and F. Flemisch. 2019. Self-determined nudging: A system concept for human-machine interaction. *Cognition, Technology and Work* 21 (2019), 621–630. https://doi.org/10.1007/s10111-019-00589-7
[59] Christopher D. Wickens and Jason S. McCarley. 2008. *Applied attention theory.* CRC Press.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009