Building trust in autonomous vehicles: Role of virtual reality driving simulators in HMI design

Building Trust in Autonomous Vehicles: Role of Virtual Reality Driving Simulators in HMI Design

Lia Morra, *Senior Member, IEEE,* Fabrizio Lamberti, *Senior Member, IEEE,* F. Gabriele Pratticó, Salvatore La Rosa, Paolo Montuschi, *Fellow, IEEE*

***Abstract*—The investigation of factors contributing at making humans trust Autonomous Vehicles (AVs) will play a fundamental role in the adoption of such technology. The user's ability to form a mental model of the AV, which is crucial to establish trust, depends on effective user-vehicle communication; thus, the importance of Human-Machine Interaction (HMI) is poised to increase. In this work, we propose a methodology to validate the user experience in AVs based on continuous, objective information gathered from physiological signals, while the user is immersed in a Virtual Reality-based driving simulation. We applied this methodology to the design of a head-up display interface delivering visual cues about the vehicle' sensory and planning systems. Through this approach, we obtained qualitative and quantitative evidence that a complete picture of the vehicle's surrounding, despite the higher cognitive load, is conducive to a less stressful experience. Moreover, after having been exposed to a more informative interface, users involved in the study were also more willing to test a real AV. The proposed methodology could be extended by adjusting the simulation environment, the HMI and/or the vehicle's Artificial Intelligence modules to dig into other aspects of the user experience.**

***Index Terms*—autonomous vehicles, human-machine interaction, driving simulator, user experience, virtual reality.**

I. Introduction

MOST research efforts in the context of intelligent vehicles (IVs) have been directed to improving safety and effectiveness of vehicle's control (autonomy) and vehicle-to-vehicle coordination (connected vehicles) [1]. To fully reap the benefits of autonomous driving (AD) systems, humans, both drivers/passengers and pedestrians alike, will need to *trust* their safety and reliability. Hence, there is an emerging need to support effective and reassuring communication between humans and IVs. Passengers need to feel confident, at all times, that they have sufficient information about the state of the vehicle, its environment and perceptions as well as its planned and current behavior; even more, that they possess all the appropriate information and means to take over all the aspects regarding the operation of the vehicle in due time, when needed, in a safe and appropriate manner.

Despite playing a crucial role in the uptake of any system based on autonomous agents, including autonomous vehicles (AVs), trust between humans and machines is generally hard to establish. According to a 2017 survey by the Pew Research Center on "Automation in everyday life", over half (56%) of the Americans who were interviewed said they would not want to ride in a driverless vehicle if given the opportunity [2].

However, preliminary experiments in the literature on partially autonomous driving scenarios show that these negative emotions can be reduced by adopting Human-Machine Interaction (HMI) designs that provide feedback about how the car is acting (what automated activity it is undertaking) and the reasons why the car is acting that way [3].

The role of HMI in IVs is thus profound and, for this reason, user experience (UX) should be taken into large account at any stage of the development process. By establishing a collaborative relationship between drivers/passengers and vehicles, HMI can positively affect the acceptance as well as the technological advancement of AD solutions.

Unfortunately, the application of consolidated approaches for UX design and evaluation to AD systems is not straightforward. For instance, focusing on the quantitative assessment of a particular user interface design, techniques that measure driver's performance in specific driving tasks could not be easily reused when, due to the specific level of automation, there are no more drivers but passengers. Similarly, post-experience questionnaires (alone) could be no more appropriate when feedback to be collected concerns the huge amount of aspects that may contribute to the perceived level of trust. Even driving simulators that are used today for developing vehicle's intelligent behaviors may not be directly applied to UX studies, as focus would have to be shifted, e.g., on the vehicle's interior and on interaction with it, rather than on the fidelity of external factors affecting its decisions (traffic, presence of pedestrians, etc.).

By moving from the above considerations, in this paper we present a methodology that is meant to support the study of HMI with IVs, and we show its helpfulness in the evaluation of the passengers' level of trust by considering the design of a possible interface for AD systems.

The devised methodology relies on a simulation platform based on immersive Virtual Reality (VR), which was developed by grounding on an existing driving simulator. Although, in principle, the technology is applicable to many scenarios, from unassisted to fully autonomous systems, we focused on L4 and L5 automation levels, as they represent the configurations for which characterizing the passenger's experience from the point of view of comfort and trust is more challenging. We therefore created a virtual AD system that allows users to experience a simulated ride in a virtual urban environment,

The authors are with the GRAINS – GRAphics And INtelligent Systems group at the Dipartimento di Automatica e Informatica of Politecnico di Torino, 10129 Torino, Italy. e-mail: (see http://grains.polito.it/people.php).

Manuscript received XXXX XX, XXXX; revised XXXX XX, XXXX.

facing a number of different situations.

For the assessment of the UX, we consider both cognitive and affective factors, by integrating feedback based on subjective post-experience questionnaires with continuous, objective information gathered from physiological signals. In particular, in this paper we focused on stress level measurements to investigate the perceived degree of safety and "connection" with the vehicle. Notwithstanding, the proposed methodology has been designed in a way to support later extensions for the detection of other emotional states. It is worth observing that, thanks to its immersive nature, VR allows to measure the latter state much more realistically than other traditional simulation scenarios [4].

With the aim to evaluate the suitability of the proposed approach, the methodology was applied to the design of a head-up display (HUD)-based interface for AVs that provides visual cues about the vehicle' sensory and planning systems. As said, providing information about how and why the car is acting is crucial to elicit trust in AVs, but little experimental evidence is available to determine how such information is best presented to the passengers [5], [6]. By applying our approach to the above scenario, we obtained qualitative and quantitative proofs that a complete picture of the vehicle's surrounding, despite the higher cognitive load, is conducive to a less stressful experience. Moreover, after having been exposed to an interface delivering a higher information content, users involved in the study were also more willing to test a real AD system.

Besides offering interesting insights that may drive future HMI designs, the results confirm the effectiveness of the proposed methodology in digging into a use case that well represents possible facets of the UX which could be investigated through the experimented techniques.

II. Background and Related Work

A. HMI in Partially and Fully Automated Vehicles

Establishing trust is important in order for users to accept, and even rely on, automated systems. Mcknight & Chervany [7] have identified three constructs necessary to increase trust: ability, benevolence, and integrity. When the trustee is an autonomous system, these factors translate in the system's *performance* and skillful execution, into the sharing of a common *purpose* with the user, and into the implementation of a reliable and consistent *process*. Trust is thus established through direct observation of the system's behavior and its underlying mechanisms. Lee & See observed that "Trust that is based on an understanding of the motives of the agent will be less fragile than trust that is based on the principle of reliability of the agent" [8]. In the context of AD systems, HMI plays a fundamental role in this respect, by providing information about the vehicle's performance. In fact, partially automated vehicles on the market allow the driver to monitor the status of the car's components. User interfaces are designed to increase the perceived ability of the system and to support predictability, thus inducing trust.

In recent years, a study by Ekman and colleagues provided a systematic review of HMI design principles that promote trust in AD systems [9]. The authors distinguish a *learning phase*, that starts with the first interaction and lasts until the user is familiar with the AD systems, from the *performance phase*, which takes into account a long-term use perspective. During a testing simulation, it can be argued that the learning phase is most important, although its specific duration differs on an individual basis. In the performance phase, trust is mainly based on the performance and dependability of the system, and is fairly stable unless an error or unexpected event occurs; in the learning phase, it is the user's ability to form a mental model of the AD systems that is crucial to form a trust bond.

Hence, in this work we focused our attention specifically on the four factors that, according to [9], are more relevant for the learning phase: *mental model*, the ability to form an approximate representation of the AD system's skills and functions; the system's proneness to be perceived as an *expert/reputable* agent; the possibility to provide continuous *feedback* to the user, ideally addressing two or more senses; finally, the provision of *how and why information* regarding upcoming actions. In this context, a "how" message describes how the system solves a given task, whereas a "why" message pertains to the motivations that lead to the task itself.

A limited number of experimental studies have, so far, established that providing information to the driver/user usually increases driving performance and acceptability in partially [3], [10], [11] and fully automated driving systems [6]. For instance, Verberne et al. [10] found that Adaptive Cruise Control (ACC) systems that share the same drivers' objectives, like the adoption of a relaxed and safe driving style without sudden braking and accelerations, while at the same time providing information to the user, are considered more reliable and acceptable. Koo et al. [3] explored the effect of providing "how" and "why" information in the context of an auto-braking system. Providing both the information types resulted in the safest driving behavior, at the expense, however, of a high cognitive load and decreased acceptability. Drivers preferred receiving only "why" information, whereas the "how" information was often perceived as redundant. The interfaces considered in these studies were very simple compared to the technical possibilities of current user interfaces: they consisted of brief verbal messages, with no visual cues [3], or included only information on the position of obstacles [6].

It is important to consider not only which information is provided to the users, but also how it is conveyed. The visual mode is the primary and most widely used among the vehicle interfaces, and represents the most consistent communication channel. In-vehicle display devices can be grouped in three categories: *head-down displays* (HDDs), *head-up displays* (HUDs), and *head-mounted displays* (HMDs). HDDs offer the advantage of not blocking the view of the real world for the users, who, however, find themselves distracting from the road. HUDs make it possible to take advantage of the necessary information while keeping an eye on the external environment, but pose significant construction challenges. HMDs share the advantages of HUDs, but only a few devices are available on the market, which suffer from some usability issues (especially for in-vehicle applications).

Studies have consistently shown that HUDs result in a better

TABLE I
INFORMATION DISPLAYED BY COMMERCIAL HUD CONCEPTS AND DEMONSTRATION VIDEOS.

AR-HUD	Displayed information
Continental AR-HUD Concept	Lane Departure Warning System (LDWS), Assisted Navigation, Adaptive Cruise Control (ACC)
Hyudai AR-HUD Concept	Traffic lights, Assisted Navigation, LKA, ACC
PSA group AR-HUD concept	Assisted Navigation, LKA, ACC, Pedestrians, Approaching obstacle warning
Daqri AR-HUD Concept	Assisted Navigation, Lane Keep Assistance (LKA), Lane Control, ACC, Approaching obstacle warning, Children crossing, Pedestrians
WayRay Holographic AR Display concept	Assisted Navigation
Waymo Demo video	Traffic signs, Cars, Pedestrians, Cyclists (bounding boxes and colored overlays with distance and speed information), Motion Prediction, Assisted Navigation
NVIDIA Drive AGX	Traffic signs, Cars, Traffic lights, Lanes, Pedestrians, Cyclists (bounding boxes with distance information), Motion Prediction, Lane separation lines, Route planning data

driving experience and performance than HDDs, leading to shorter reaction times [12], decreased cognitive load [13], and fewer driving errors [5], [14]; HUDs are also preferred by users against both HDDs and HMDs [5]. Augmented Reality HUDs (AR-HUDs) have been found especially effective in increasing the driver's intuitive cognition [15] and promoting a safer and more effective driving behavior, particularly in demanding driving situations [16], [17].

Given the technical difficulties in realizing AR-HUDs, current displays often come in the form of prototypes, concepts or demonstration videos. Examples in the literature often focus on specific aspects of the driving experience, such as driver assistance (DA) [5] and obstacle detection [6]. Many commercial prototypes focus on partially automated systems that extend current DA solutions, whereas Waymo and NVIDIA are more directly focused on L4 and L5 automation.

Information displayed by the main commercial solutions is reported in Table I. A tendency to adopt a common set of symbols and metaphors can be observed among vendors. For instance, information related to ACC and Lane Keep Assistance functionalities, such as the current lane, speed, and the position and speed of preceding cars are displayed by Continental, Hyundai, PSA, and Daqri. Waymo and NVIDIA include richer information on both the path planning and the sensory capabilities of the vehicle. Through bounding boxes (i.e., parallelepipeds enclosing detected objects), colored overlays and other elements, all the factors involved in driving are highlighted. In addition, navigation information is added not only for the user's vehicle, but also related to other cars, pedestrians or cyclists through motion prediction.

B. Measuring User Experience in Driving Simulators

Researchers have for long time relied on driving simulators to cope with difficulties and risks associated with field testing [18]. In recent years, VR simulators have elicited a lot of interest thanks to their immersive nature [5], [13], [19], [20]. Most studies investigating different aspects of driving in simulated scenarios, including HMI design [5], [21], rely on drivers' behavior and performance as a proxy for their emotional and cognitive status [3], [12], [22]. Experimental measures include standardized questionnaires as well as indicators such as driving speed, lane keeping, braking patterns, etc., for which absolute or relative validity has been generally established [22]. However, in AD systems, humans are expected to take progressively less part in driving, which makes behavioral assessment less relevant.

Physiological signals are increasingly used to measure users' affective and cognitive states in engineering in general [23]. The activity of the autonomic nervous system, which regulates affective states, can be captured non-invasively through signals such as Heart Rate (HR) and Electrocardiography (ECG), Electromyography (EMG), Respiratory Rate, and Galvanic Skin Response (GSR). In the last years, researchers also investigated their use combined with traditional or immersive driving simulators [4], [22], [24].

In particular, the relative validity of physiological signals for traditional driving simulators is supported by several studies, albeit available data is less abundant than for driving performance [4], [22], [25]. For instance, risk perception was found to be highly correlated with changes in GSR [22]. Comparison between on-road and simulated driving conditions established the relative validity for mean HR and mean oxygen consumption, although HR values observed in real driving conditions were higher, probably due to the increased stress associated with driving on a real road [25]. In a pilot study, Eudeave and colleagues found that the physiological response in an immersive VR environment is stronger than in a traditional driving simulator [4].

Recording of physiological signals have also been exploited in real-life driving conditions to characterize drivers' performance and experience, from measuring stress levels to detecting drivers' drowsiness [26], [27]. Of particular interest is the study of Healey and colleagues on driving-related stress [26]. ECG, EMG and GSR were recorded while drivers followed a set route; driving sessions were videotaped and visually inspected for observable stress-induced actions, such as head turning, to be used as reference standard. Collected signals allowed the authors to distinguish different levels of stress with high accuracy (over 97% across multiple drivers); GSR and HR metrics were most closely correlated with drivers' stress level. Again, studies have been conducted, so far, from the point of view of an active driver, leaving the question open on whether stress-induced changes can be equally and as effectively observed in passengers.

III. PROPOSED METHODOLOGY

A. Overview

As discussed in Section II-A, trust in automated systems can be achieved from direct observation of system's behavior, coupled with an understanding of the underlying mechanisms. To this aim, as depicted in Fig. 1, the devised methodology relies on a VR-based AV simulator. Simulation allows the user to get immersed in repeatable scenarios including a

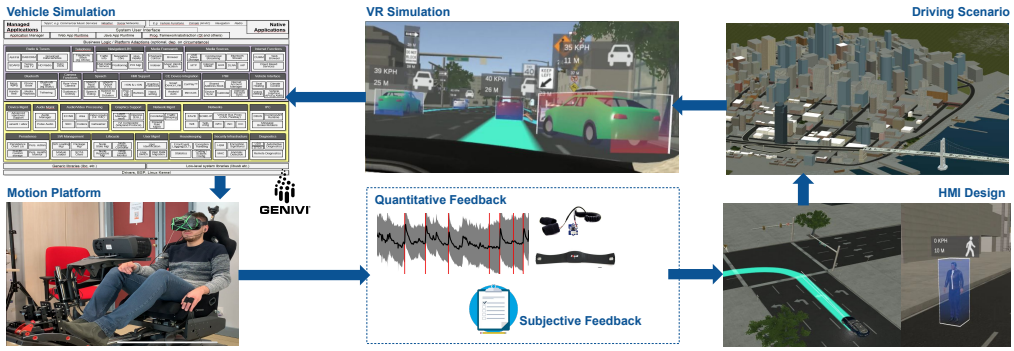


Fig. 1. Proposed methodology: exemplification on the task of HMI design. A defined AV driving scenario is simulated in immersive VR. Vehicle simulation is based on the open source GENIVI platform. The simulator is integrated with a motion platform to further foster immersion. User's feedback is collected through both subjective, offline questionnaires, and objective, real-time physiologic measurements reflecting cognitive and affective states.

variety of both ordinary and emotional-intensive events. User is provided with insights on the autonomous system's behavior by means of a virtual AR-HUD combined with additional audio cues. In this way, we postulate that the user can form an adequate mental model of the AD system. Assessment is performed by collecting feedback from the user in the form of subjective (questionnaire-based) ratings and objective (physiological signal-based) measurements.

B. Technology and Setup

The VR system is implemented using the HTC Vive ecosystem (https://www.vive.com/eu/product/), by HTC Corporation, Taiwan. The Vive VR Headset features a resolution of 1080×1200 pixels per eye spanning a horizontal 110° FOV at 90Hz. The native positional tracking leverages the IR lasers emitted from the Vive Base stations (built upon the Valve's Lighthouse technology) which, combined with headset's built-in sensors, enables a 6 DOF *outside-in* tracking of the user's head.

With the aim to foster immersion through the simulation of the motion stimuli that a driver or passenger would experience on a real vehicle, an inertial motion platform is used. The platform exploited in this work is the Atomic A3 Racing, designed by Atomic Motion Systems, which supports 2 DOFs (yaw and pitch) motion simulation. To simulate the user's perceived accelerations, the so-called *tilt coordination* motion simulation strategy [7] was implemented. In short, this technique works by imitating the perceived acceleration via decomposition of the gravity acceleration vector, obtained through a coherent rotation of the platform. A motion compensation needs to be applied to the VR coordinate system (which is centered in the headset, i.e., in the user's viewpoint), based on current platform's rotation. To this purpose, a Vive Tracker was mounted on the seat and tracked together with the headset.

Finally, since it has been proved that letting the user see his or her hands in the virtual environment increases the sense of presence [28], a virtual replica of the user's hands including articulated fingers is created by tracking them using a Leap Motion Controller device attached to the headset.

C. AV Driving Simulator

The vehicle simulator implemented is based on the open source Simulator Vehicle project by the GENIVI Alliance [29] (in the following simply referred to as GENIVI, for brevity). GENIVI was selected among several possible alternatives for multiple reasons: it was originally created to support HMI design; it allows, by design, the addition of new features; it already provides modules for intelligent traffic simulation; it includes a basic *auto drive* functionality for the user's vehicle; it provides a few driving scenes and vehicles with their own rigid body physics-based controller.

The main activities carried out to adapt GENIVI to the purposes of this work involved: porting of available features to VR; integration of the motion platform; implementation of a custom AD controller. The latter activity was considered as necessary since, in a preliminary study, the built-in controller was judged not realistic enough, especially when dealing with complex, unpredictable events (e.g., sudden pedestrian crossing, etc.).

1) VR porting: Implementing the support for VR was facilitated by the fact that GENIVI is based on the Unity game engine, which natively allows for the creation of VR applications for the HTC Vive. Our implementation allows to virtual accommodate the user to any seat of the virtual vehicle. Built-in vehicles, namely a Land Rover L405 and a Jaguar XJ, are designed for a non-immersive simulation. Hence, a new vehicle was created with VR-based interaction in mind, i.e., by focusing on visual fidelity of the vehicle's interior. Finally, support for users' virtual hands was added.

2) Motion platform integration: to integrate the motion platform, an additional software module was developed. The module receives in input the acceleration values calculated in the seat's tracked point by the physics simulation engine and outputs it to the proprietary platform's driver (AMS Symphinity), which remaps them to coherent tilt and pitch angles and consequently applies them to the platform. Other motion platforms may be integrated in a similar way.

3) AV simulation: within this work, our aim was to provide a methodology to study the considered domain using simulated VR-based scenarios, accompanied by suitable measurement tools, rather than contribute to the advancement of the state of

the art of AVs' control sub-systems. Basic AD functionality available in GENIVI was therefore extended to make it cope with situations of interest. Attention was focused on reproducibility, while preserving simplicity. More sophisticated implementations, leveraging, e.g., data provided by vehicle's virtual sensors could nonetheless by integrated in the future.

Our implementation takes advantage of the native trajectory system, which is used in GENIVI to manage the traffic. Paths to follow are embedded in the scene description using a complex network of waypoints. The developed AD system relies on it to feed a PID-based controller, which is in charge of driving the vehicle by making it accelerate, brake, and steer. Differently than with the other cars in the traffic, the AD system is affected by the full set of accurate, rigid body physics simulation variables. The PID was fine-tuned in closed loop using manual parameter adjustment targeting a maximum overshooting of 5% at step response, in order to achieve a comfortable and realistic behavior. To this aim, control commands shaping and auxiliary waypoints were also used. Although different and far more sophisticated approaches could be investigated in the future (e.g., [30]), the selected control system proved to meet the simplicity-effectiveness trade-off required to cope with the issues tackled in this work. Appropriateness of the pursued approach was also confirmed by subjective observations concerning simulation quality (Section IV-E and Supplemental material).

The same approach was pursued also to bind specific vehicle's reactions to the pre-programmed events. Obstacle avoidance is handled by a dedicated logic, which also takes into account trajectory replanning when the obstacle cannot be avoided by simply adjusting vehicle's speed. For moving obstacles, replanning takes into account predicted motion. Further details on the implementation can be found in [31].

D. HUD Design

Based on the principles discussed in Section II-A, an in-vehicle user interface should continuously provide feedback addressing, whenever possible multiple senses, highlighting "why" information that explains the vehicle's choices, and adopting a pleasant and effective communication style that presents the system as a skilled/reputable driver. These elements, while important in general, are particularly relevant in the initial learning phase, where the user is still unfamiliar with the AD system and needs to form an appropriate mental model of its inner mechanisms [9]. As it will be illustrated in more detail in Section IV, subjects who participated in our study were never exposed to a real AD system. An AR-HUD was therefore designed, as it was found in the literature to be the most effective interface under the considered conditions.

It was deemed as important to ensure that visual cues displayed by the AR-HUD are consistent with information conveyed by commercial DA products, as users are mostly familiar with it. However, it was regarded as crucial to provide also information that illustrate the vehicle's sensory capabilities and hence, improve the user's situational awareness. Finally, given this work's focus on L4 and L5 AD systems, information about the vehicle's planning functionalities needed to be delivered as well.

Design was based on the features reported in Table I. The HUD is capable to display information about all the relevant elements in the surrounding environment, including both static objects (trees, lighting poles, parked cars, traffic lights, road signs, etc.) and dynamic objects (pedestrians, animals or other cars). These elements are provided together with distance information in meters from the vehicle, absolute speed when available, and a visual warning status indicator. Lane keeping and navigation cues for the user's vehicle and other cars (assuming that they are connected) are also considered. The color of each car is randomly assigned by GENIVI.

Objects of interests are identified by means of a bounding box. This metaphor, previously validated in the literature [17], is adopted by commercial players such as Waymo and NVIDIA. In our implementation, each bounding box has a white outline and is associated with a label and an icon identifying the detected object, thus satisfying the usability principle which suggests that the adopted representation must be simple and intuitively understandable by the user [32]; the use of familiar cues, such as icons, also reduces the cognitive load in the presence of a large amount of information [33]. Bounding boxes are automatically generated in VR knowing the position, size and pose of all objects in the scene. Technically, this was implemented in Unity by associating to all objects with a *Collider* component a visible colored material. The *Colliders* are not visible by default in the rendering step because just define the bounding volume of an object for the purposes of identifying physical collisions through the physics engine. Labels always face the vehicle and are, therefore, readable by the user.

In order to determine which situations constitute a potential danger, we relied on the definition from the ISO 15623 standard on "Forward vehicle collision warning systems" [34], counting on a previous study by Sebastian et al [35]. A mathematical model is used to determine potential collisions based on the trajectory, speed and acceleration of the vehicle, as well as that of potential obstacles (e.g., the preceding car). Once the possibility of a collision has been established, a safety distance is calculated, which depends on the speed of the vehicle and the reaction time of the driver, which was estimated based on the study in [36]. The distance between the vehicle and the estimated collision point is therefore measured: if this distance is less than the safety distance, the passenger needs to be warned of the potential danger. We therefore defined a *hazard index*, ranging from 0 to 1 and calculated as the ratio between the distance from the obstacle and the warning distance defined in [35].

The objects' warning status is presented through both visual and auditory cues. In the literature, the AR-based DA system in [37] adopted an intuitive color code in which the severity of the danger of an obstacle detected on the road is shown by means of a color code that starts from the green (safety) and extends up to the red (maximum state of danger). This color coding is consistent with systems reviewed in Section II-A. Therefore, we decided to color-code the hazard index with a green to red gradient and use it to visually represent the warning status of the detected object by controlling the transparency color of the bounding box. The color-code value is computed using a

perception-based equation, where the hazard index is used as exponential factor.

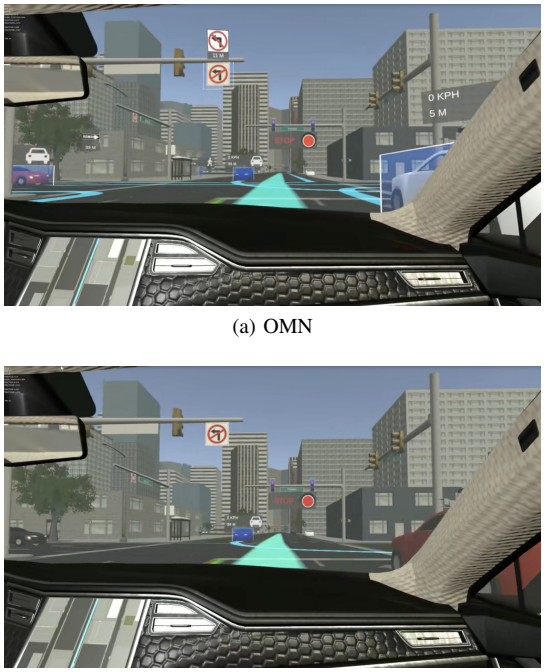
To signal potential dangers, the label associated with the bounding box flashes as well, so as to direct the user's attention towards the obstacle. Flashing is used in DA systems by various vendors [38], [39]. It is important to underline that, through the flashing information, the vehicle communicates to the user why it is about to perform a specific action [9]. The flashing visual cue, at a lower frequency, is also used to notify the user of a road sign or traffic light. Flashing occurs when a traffic light changes or when a new road sign is recognized. The lower frequency reduces the sense of alarm and, hence, allows the user to distinguish normal driving operations from high-risk situations.

An immediate danger is also marked by a sound alert [34], [40]. This is consistent with current DA systems, which produce audible warnings, e.g., in emergency braking conditions [39]. A more pleasant sound is played when road signs are detected, to capture the user's attention in an unalarming way.

Two variants of the HUD were designed, which in the following are referred to as *omni-comprehensive* (OMN) and *selective* (SEL).

1) Omni-comprehensive HUD: in the OMN variant, we show information about all dynamic elements (cars and pedestrians) within a "detection" diameter which is set to 150 meters. This threshold was firstly motivated by practical reasons: virtual objects beyond this distance would be too small to be appreciated considering the resolution of the display in the VR headset. This distance is also compatible with the equipment of current AD prototypes and the detection range of LiDAR systems. Road signs and traffic lights are always shown in the interface, except for those that regulate road sections different than the one the vehicle is currently on. Furthermore, it was decided to exclude from the display the information about static objects such as trees, parked cars, lighting poles, etc. unless they become dangerous. This exclusion is motivated by the principle of cognitive load, according to which an interface should be easily understandable by the user, simple and intuitive, as it avoids excessive cluttering [32].

2) Selective HUD: in the SEL variant, only information that is deemed of specific interest to the user is displayed. The guiding principle was to select information that pertains to those elements of the environment that, at any given point, affect the behavior of the AD system. Let us consider road signs: the vehicle detects all road signs within the diameter of interest, but not all of them are necessarily useful at the time. For example, in the presence of a pedestrian crossing sign and a speed limit sign, the vehicle may decide not to show any information on the former sign, based on the fact that, at the moment, there is no pedestrian intending to cross; the latter sign may force the vehicle to slow down and, thus, it would be highlighted in the interface. More specifically, in the SEL variant only cars that precede the current vehicle or, more generally, that intersect its current trajectory, are highlighted with a bounding box. Pedestrians and other static or moving objects are identified only if and when they become dangerous, i.e., when a collision becomes possible. Navigation lines for other cars are only displayed when assessed by the vehicle (e.g., at intersections to determine priority). Traffic lights information, as well as tracing of the vehicle's navigation line and the road center line, are unchanged in this variant. A comparison of the two interfaces is provided in Fig. 2.



(a) OMN

(b) SEL

Fig. 2. Comparison between the OMN (a) and SEL (b) AR-HUD interfaces.

E. Simulated Scenario

In order to create a relationship of trust between a user and an AD system, the latter must show its ability in dealing with different driving scenarios [8]. Our simulated scenario is constructed to include a variety of different situations, both ordinary and challenging, that may occur in an urban setting. The urban setting is, in general, considered the most difficult to manage by AVs [41]. In fact, current L2 and L3 automated systems are mostly restricted to motorways and extra-urban routes, and the biggest challenge in the development of L4 and L5 systems is represented precisely by urban areas where significantly more factors are at play, driving conditions are far less predictable, and the presence of pedestrians amplifies the perceived risk.

Compared to real-life driving, simulators offer the distinctive advantage of creating repeatable scenarios, where most experimental factors can be easily controlled. Therefore, it is possible to study and compare subjects' reaction to individual events, whereas in real-life driving experiences one would be mostly restricted to consider overall measures [25], [26].

The simulation was created starting from one of the scenes included in the GENIVI platform, representing a miniature version of the city of San Francisco. As said, despite the basic auto drive functionality, GENIVI is natively meant to support mostly first-person driving experiences, with random traffic patterns, no pedestrians and no intentionally hazardous situations. By leveraging the developed AD capabilities and

the integrated waypoint system, different situations were embedded in the simulated scenario in order to showcase different AD abilities and elicit changes in the subjects' affective state. Considered abilities include: interacting with traffic and especially with other cars, e.g., maintaining safety distance, overtaking, etc.; handling road signs and traffic lights; avoiding obstacles and dealing with other potentially dangerous situations, including those where other cars or pedestrians do not behave correctly.

The subject is seated in the passenger's position (front right). The experience begins in an area with a relatively simple environment and little traffic, to allow the user to familiarize with the AD system. Then, the environment becomes populated by cars and pedestrians: the subject becomes familiar with the HUD and the way information is conveyed by the vehicle. Afterwards, riskier situations occur in which the vehicle can show its decision-making skills. In [8], it was observed that "If trust is primarily based on rules that characterize performance during normal situations, then abnormal situations might lead to the collapse of trust". This strengthens the importance of including driving situations that, while less likely, may pose significant challenges for an AD system. To simulate a typical urban context, such situations were spaced throughout the simulation and alternated with ordinary ones, as illustrated in Fig. 3. After every risky situation, the car stops for few seconds, to ensure that the subject has enough time to understand what happened and to reflect on how the car handled that situation. Considering the time for letting subjects get acquainted with the system as well as the time required to achieve a suitable distribution of situations, the duration of the simulated scenario was set to 12 minutes.

Simulated events include the sudden crossing of a dog (Dog), a child on the sidewalk throwing a ball on the street (Ball), scooters and cars that split lanes while driving (Scooter, Car1 and Car2), as well as pedestrians crossing the street (Man1 and Man2). Illustrative frames are reported in Fig. 3. The Dog event corresponds to a highly hazardous situation, in which the vehicle is forced not only to slow down, but also to steer in the opposite direction to avoid a collision. The same happens in the Ball and in the Man2 events (in the latter case, a pedestrian crosses outside of a designated crosswalk while the car is at full speed). Man1 is a less risky situation, as the car is approaching a red light and is already braking when the pedestrian starts crossing. In the Scooter event, the vehicle slows down as the preceding car turns right; in the meanwhile, a scooter enters the lane from the left. The situation is not particularly dangerous, as the vehicle was already reducing its speed to deal with traffic jam; however, from the viewpoint of vehicle-to-human communication, this interaction is complicated as it involves several vehicles. In the Car1 event, a car suddenly changes its lane when approaching road construction (which is poorly visible), forcing the vehicle to quickly reduce its speed to avoid a collision. The Car2 event is even riskier, as another car driving on the intersecting road does not stop at a red traffic light and instead passes at full speed, forcing the vehicle to brake very abruptly.

Two videos showing the simulated scenario with the OMN and the SEL interfaces are available at http://tiny.cc/p4v16y.

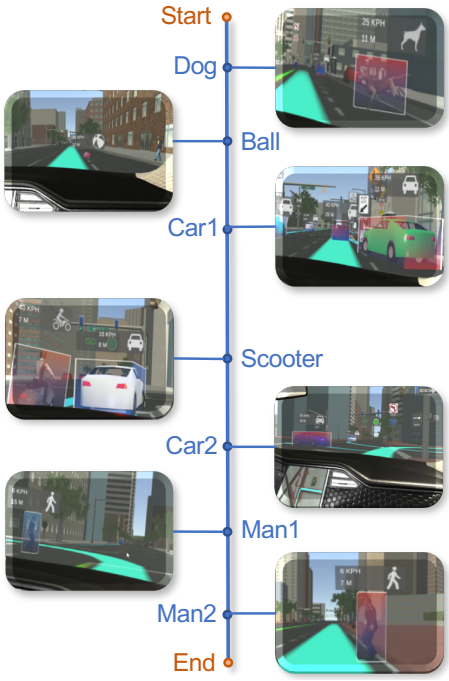


Fig. 3. Timeline of the test scenario with simulated events.

F. Galvanic Skin Response

As discussed in Section II-B, physiological signals related to the activity of the autonomic nervous system can provide non-invasive information about the user's affective state. While, in principle, a combination of different signals can be used, for the sake of simplicity in this work we focus on the GSR signal, which was found to effectively detect stress in both simulated and real-life driving [22], [26]. Furthermore, it is easily measured with a simple sensor placed on the fingers [23]. GSR is mostly sensitive to the dimension of *arousal*, going from sleepiness to excitement or stress [42]; it leaves open whether the arousal change is of a positive or negative nature (the *valence* dimension), which nonetheless in our specific case can be derived from the context.

The GSR can be decomposed into a slowly changing tonic component, the Skin Conduction Level (SCL), and an impulsive phasic component, the Skin Conductance Response (SCR) [42]. While the SCL reflects the overall emotional state as well as habituation to the environment, the SCR measures activation in response to a stimulus, e.g., a potentially stressful event occurring in the simulated scenario. The magnitude of the response should correlate to the perceived threat. This phenomenon was previously validated in other types of VR environments [43], with an observable effect on the GSR signal even after multiple exposures.

1) Signal processing: the SCR data was extracted using a 3rd order Butterworth band-pass filter ranging from 0.16 Hz to 2.1 Hz [27], [42]. Normalization is required to account for the intrinsic inter-individual differences in skin conductance [23]. The most common choices are *z-score* standardization, in which the signal is divided by the standard deviation after subtracting the mean, and *min-max* normalization, in which the signal is normalized between 0 and 1. We found that the min-

max scaled signal was most useful for visualization purposes and trend analysis, whereas z-score normalization could be used for the final feature extraction as we observed less inter-subject variability. Considering the fact that, typically, SCR peaks appear between 1 and 5 seconds from the stimulus's onset and last for about 10 seconds, we extracted the SCR waveform for a time window of ±10s centered on each event [44], [45]. Within each window, all samples are divided by the initial signal value, to focus on relative changes.

2) Feature extraction: for each event time window, a set of features is extracted [44]. Let $\hat{GSR}(k,j,i)$ be the z-score standardized data value for subject k, event j and sample at time i, $SCR(k,j,i)$ the corresponding filtered signal representing the skin conductance response, and L the total number of samples per time window. Features extracted include the mean GSR (Eq. 1), the accumulated GSR (Eq. 2), the max GSR (Eq. 3, and the Peak to Peak distance in SCR (Eq. 4). Each feature is calculated on the 10s before (Pre) and the 10s after ($Post$) the test event, and the difference (Δ) is used as the final measure.

$$\overline{GSR}_{mean}(k,j) = \frac{\sum_{i=0}^{L} \hat{GSR}(k,j,i)}{L} \quad (1)$$

$$GSR_{Acc}(k,j) = \sum_{i=0}^{L} \hat{GSR}(k,j,i) \quad (2)$$

$$Max(k,j) = \max_i \left(\hat{GSR}(k,j,i)\right) \quad (3)$$

$$P2P(k,j) = \max_i \left(SCR(k,j,i)\right) - \min_i \left(SCR(k,j,i)\right) \quad (4)$$

G. Questionnaire

Subjective data about the experience can be collected through questionnaires. The questionnaire that we designed tackles factors affecting trust and, in general, HMI effectiveness [9]. Specific sections were included to test each of these factors. The questionnaire includes both general questions, that could be re-used across different driving scenarios, as well as questions that are more specific to HMI and to the simulated scenario. We focused our attention on those aspects that are more relevant for establishing trust in an initial learning phase, where the user gets acquainted with the system. When possible, questions were mutuated from validated tools such as Simulator Sickness Questionnaire (SSQ) [46], the Situation Awareness Rating Technique (SART) [47] and the NASA Task Load Index (NASA-TLX) [48]. Questions were organized in the following sections.

1) Health status: VR systems may induce motion sickness and other side effects: to avoid biases, health status is collected before and after the experience using the SSQ tool.

2) System competence: Inspired by standard questions for the evaluation of trust in human-robot interaction (HRI), this section evaluates the perceived system's *competence* across the range of driving situations explored in the simulation [49].

3) Reaction to test events: For each test event, the user is asked to rate four statements: *1) The situation was dangerous*, *2) The event took me by surprise*, *3) I was able to see the potential danger before it affected the vehicle's performance*, and *4) The interface provided me useful information to foresee the event*. These questions provide complementary information to the physiological signals and disentangle the effect of the specific event from the HMI.

4) Situational awareness: This section was inspired by the SART tool, focusing on dimensions (quality, quantity and familiarity) that pertain to comprehensibility. Here, quality refers to the usefulness with respect to clarifying system's intentions. Quality and quantity were evaluated for each element of the HMI, e.g., bounding boxes, navigation lines, etc.

5) Cognitive load: This section was adapted from the NASA-TLX evaluation tool.

6) Overall user experience: This section investigates general aspects regarding the mental model, and is concluded with a direct question about trust. Predisposition towards participating in an AD experience was also assessed before and after the simulation.

7) Immersion and presence: Immersion, presence and simulation fidelity were evaluated by adapting the relevant sections from the VRUSE questionnaire [50], an established technique to measure usability of VR applications.

All questions were in Italian and had to be rated on a 1–5 Likert scale. Sections *Reaction to test events* and *Situational awareness* included snapshots of the test events and the HMI elements, respectively; the questionnaire was adapted for each test group with snapshots from the specific HUD version. The complete questionnaire (SEL version) is available at https://forms.gle/CpSYZc729fho7gy86.

IV. Experiments

A. Data Acquisition

Healthy individuals (e.g. with no impairing chronic or acute illnesses at the time of the acquisition) with a valid driving license were recruited to participate in the virtual driving experiment. Participation was voluntary and no monetary compensation was provided. Study participants were randomly assigned to either the OMN or SEL HUD group. All acquisitions were performed within one week.

The test phase began for each subject with a brief explanation of the test session. Health status, demographic information and general disposition towards AD systems were collected before starting the simulation. Two baseline signals were also acquired: one minute at rest, and one minute after placing the VR headset. After the simulation, the final questionnaire was administered and the experience debriefed.

The GSR was recorded through an ad-hoc device based on the Groove GSR Sensor [51] and a Raspberry Pi 3 board. The acquisition module was implemented in Python. An external Analog to Digital Converter (MCP3008 [52]) was used to connect the output of the sensor to the board via the Raspberry's Serial Peripheral Interface (SPI). The sampling frequency was set to 256 Hz in order to separate the two components of the GSR signal [44]. Due to inter-subject variability, the GSR may saturate during the analog to digital conversion: therefore, during the initial baseline acquisition, the converter was manually calibrated by adjusting the resistor until the output fell in the 200–512 a.u. range. The sensors were applied on the fingers of the non-dominant hand, after washing the hands. Postprocessing and feature extraction was

implemented in Python 3.6.5 and the SciPy library for filtering; all calculations were performed on an HP Pavilion, Intel Core i5-3230M CPU.

B. Statistical Analysis

A two-way factorial Analysis of Variance (ANOVA) was conducted to examine the main effect of HUD as well as the interaction effect between event and HUD type on each GSR feature. A mixed design was employed with the HUD type as the between-groups factor and the event as the within-subjects factor. Post-hoc comparison between the different events and HUD types was performed applying Bonferroni correction.

Questionnaire data was analyzed separately for each group of questions. Event-related questions were analyzed using a two-way factorial ANOVA, using the same design of the GSR feature. Outcomes of the other questions were compared between the OMN and SEL groups using the Mann-Whitney U-test for categorical data. A p-value of .05 or lower was considered to indicate a statistically significant difference. Statistical analysis was performed using SPSS v20, whereas signal analysis and feature extraction were coded in Python.

C. Participants' Characteristics

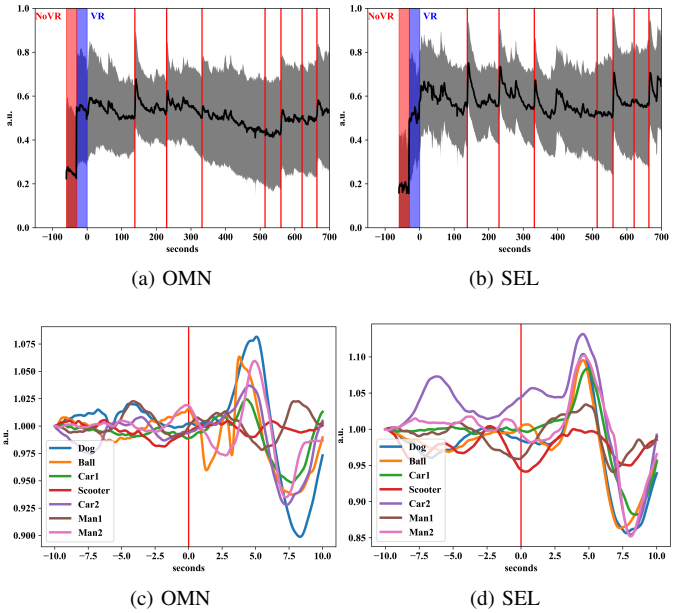
Thirty-nine subjects volunteered to participate in the study. One subject with excessive motion sickness was excluded from the data-set, as symptoms would bias the physiological response [53]. A total of 38 subjects (25 male, 13 female, mean age 23.9) were included in the analysis. GSR data was not available for 8 subjects due to failures in the recording equipment. Most of the subjects reported using VR or driving simulators "never" or "rarely" (30/38 and 34/38, respectively).

D. Quantitative Measurements

The normalized GSR signals averaged over all study subjects within each group are reported in Fig. 4(a)–(b). All subjects showed an increase in baseline GSR in VR. Moreover, a noticeable peak in the GSR occurred for most events in the test scenario. Fig. 4(c)–(d) show the mean SCR curve for each event. Each curve is extracted for a time window of ±10s centered at each event; within each window, all samples are divided by the first value to highlight changes.

From the SCR and GSR curves, different features have been extracted, as defined in Section III-F2. We here report in detail the two-way ANOVA results for the $\Delta P2P$ feature. The main effect of HUD was significant, F(1,28)=4.72, p=.039, indicating a statistically significant difference between the OMN and SEL interfaces. The main effect of event was also statistically significant, F(6,168)=13.9, p<.001. We did not find a significant interaction between HUD and event, F(6,168)=1.74, p=.115; hence, post-hoc analyses were conducted on each main effect separately.

The mean and standard error of the SCR feature for each event and for each HUD are reported in Fig. 5. At post-hoc analysis, the SEL HUD consistently showed higher emotional arousal for Car1 (p=.022), Car2 (p=.042) and Man2 (p=.041) events. For the first two events in the timeline, a positive trend could be observed (p=.181 and p=.409). For the Scooter and Man1 events, which elicit no emotional arousal, differences were not statistically significant (p=.759 and p=.990).



(a) OMN (b) SEL

(c) OMN (d) SEL

Fig. 4. Normalized raw GSR signal for OMN (a) and SEL (b) interfaces; baseline is collected prior to the experience and red lines represent test events in the simulation. Average SCR curves over all subjects in the 10s before and after the event for all the test events for OMN (c) and SEL (d) interfaces.

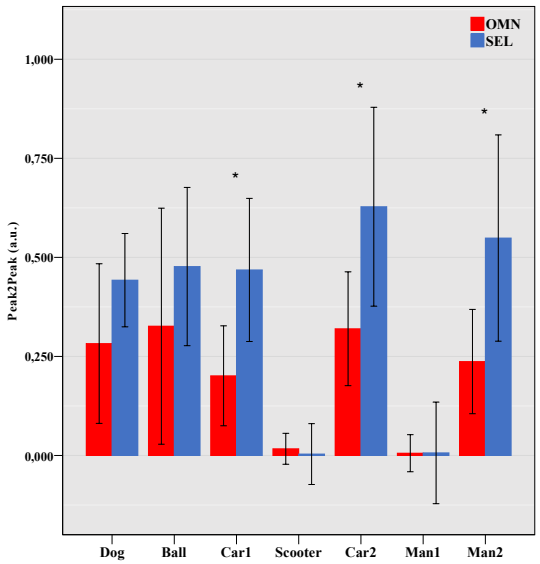


Fig. 5. $\Delta P2P$ feature, for all events, with OMN and SEL. * p-value<.05.

All GSR features showed a significant main effect of HUD: ΔMax, F(1,28)=8.53, p=.007; ΔGSR_{Mean}, F(1,28)=9.36 p=.005, and GSR_{Acc}, F(1,28)=9.02, p=.006. Likewise, a significant main effect of event was always found (p<.001), with no significant interaction between HUD and event. At post-hoc analysis, results for the ΔMax feature were comparable to $\Delta P2P$, whereas ΔGSR_{Mean} and ΔGSR_{Acc} features reported significant differences (p<.05) for Ball, Car1 and Car2 events, instead of Car, Car2 and Man2 events.

For each HUD and each test event, $\Delta P2P$ *Pre* (10s before the event) and *Post* (10s after the event) values were tested for differences using two-tailed t-tests. With the SEL HUD, all events showed a significant increase in SCR (p<.001), except for Scooter (p=.927) and Man1 (p=.920). Very similar results were obtained with the OMN HUD: the Dog (p=.014), Ball (p=.046), Car1 (p=.007), Car2 (p=.001) and Man2 (p=.003) events showed a significant effect on SCR, whereas Scooter (p=.142) and Man1 (p=.422) did not. Results for other GSR features, here omitted for brevity, were also consistent.

E. Questionnaire Results

Only one subject was excluded from this analysis due to high motion sickness; the other subjects did not report excessive symptoms (nausea rating M=1.26, SD=.54).

Subjective ratings for test events are reported in Fig. 6. Four statements were included for each test event, as detailed in Section III-G3; for the sake of clarity, only question 1 (which evaluates the risk) and question 3 (which evaluates the ability to detect the potential danger in advance) are included in the plots, as answers to questions 2 and 4 were very similar. At two-way ANOVA, the main effect of both HUD, F(1,36)=15.91, p<.001, and event, F(6,216)=54.05, p<.001, on the perceived risk (question 1) were statistically significant. Interaction between the two factors failed to reach statistical significance, F(6,216)=2.05, p=.060. Regarding the ability to identify dangerous situations in advance (question 3), the main effect of both HUD, F(1,36)=28.08, p<.001, and event, F(6,216)=14.78, p<.001, were statistically significant, without a significant interaction, F(6,216)=1.75, $p = .112$.

Since events are the same in both groups, we attribute the difference in perceived risk to the greater ability of the OMN interface to convey information about the vehicle's surroundings before critical situations occur. At post-hoc analysis, differences were statistically significant for Car1 (p=.003), Car2 (p=.017) and Man2 (p=.008) events, and a positive trend was observed for Dog (p=.134) and Ball (p=.872) events.

For each event, questionnaire ratings and GSR features values were compared by using multiple linear regression; by attempting to predict the average GSR outcome ($\Delta P2P$) from the average questionnaire ratings, we can desume the degree of similarity between the two measurements. A per-subject analysis was not attempted, given the limited sample size. A statistically significant regression equation was found, F(4,9)=14.34, p=.0007, with an adjusted R^2 of 0.804, which indicates that roughly 50% of the variance of the GSR can be explained by the questionnaires. Individual factors failed to reach statistical significance, but the strongest trends were observed for the perceived level of risk (coefficient 0.111, p=.29) and the element of surprise (coefficient 0.112, p=.29), which are presented in the scatter plots of Fig. 7.

Subjects generally found the vehicle's driving skills adequate (SEL M=4.53, SD=0.61, OMN (M=4.68, SD=0.48, p=.556). In the SEL group, subjects reported more often that the vehicle faced difficulties with unexpected changes in the environment (SEL M=1.68, SD=0.75 and OMN M=1.21, SD=0.42, p=.41); such differences can only be attributed to the HUD, considering that the vehicle's behavior was exactly the same in both experiences.

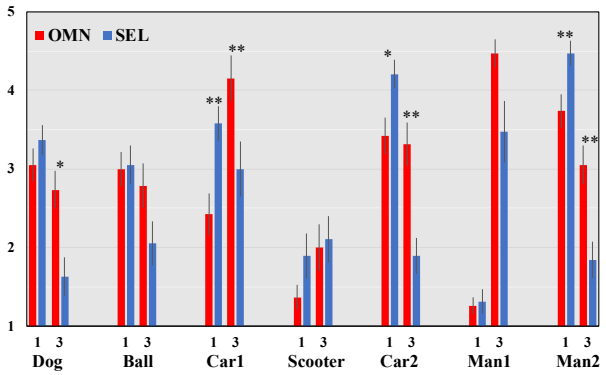
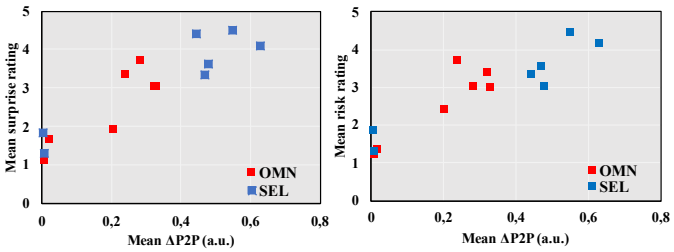


Fig. 6. Subjective measurements for questionnaire section *Reaction to test events*. Label 1 refers to the question which evaluates the risk perception, on a scale from 1 (low risk) to 5 (high risk); label 3 refers to the question that evaluates if and how the individual previously noticed the dangerous situation, on a scale from 1 (not previously noticed) to 5 (previously noticed). Each test event is considered separately. $^*p$-value <.05, $^{**}p$-value <.01.



(a) (b)

Fig. 7. Comparison of subjective vs. objective ratings. For subjective measurements, the average perceived risk (a) and the average surprise rating (b) are reported (where the latter refers to the extent to which the user was taken by surprise by the event). The mean $\Delta P2P$ feature is reported as objective rating. Each data point corresponds to a specific test event.

Displaying more information may result in an excessive cognitive load. Indeed, subjects in the OMN group more often rated the amount of information provided by the interface as excessive (OMN M=2.1, SD=0.229, SEL M=1.05, SD=0.809, p<.001), whereas comprehensibility was rated adequate for both the interfaces (p=.908). On average, the UX was satisfactory for both the interfaces, and the information provided by the HUD was considered useful (SEL M=4.16, SD=0.69, OMN M=4.84, SD=0.38, p=.001). Participants in the OMN group reported that the information was more useful in order to understand why the vehicle made a decision (SEL M=4.26, SD=1.05, OMN M=4.84, SD=0.38, p=.055) and to feel in general at ease (SEL M=3.79, SD=1.08, OMN M=4.68, SD=.59, p=.003), as well as that the vehicle seemed to have greater control on the external environment (SEL M=3.79, SD=1.08, OMN M=4.84, SD=0.38, p< .001). Overall, the OMN HUD was more helpful in anticipating potential dangers (SEL M=2.42, SD=0.61, OMN M=4.10, SD=0.57, p<.001). Subjects reported a high sense of immersion (M=4.50, SD=0.73) and presence (M=4.37, SD=0.59), with no significant difference between the two groups.

Finally, users in the OMN group were better disposed

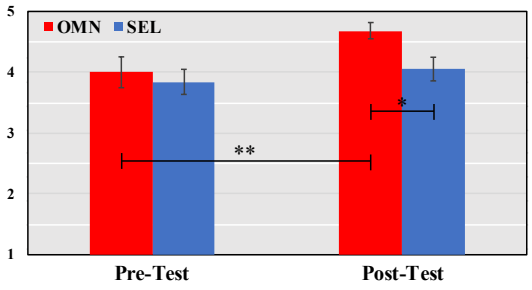


Fig. 8. Disposition towards participating in a real AD experience. On the left, the pre-test answer for the SEL and OMN interfaces; on the right, the post-test answer. Scale from 1 (absolutely negative) to 5 (absolutely positive). Mann-Whitney U-tests between pre- and post-test answers are shown, $^{*}p$-value$<.05$, $^{**}p$-value <0.01.

towards participating in a real AD experience (OMN M=4.68, SD=0.58, SEL M=4.05, SD=0.85, p=.012). As shown in Fig. 8, prior to the experiment all participants were mildly optimistic, but after experiencing the OMN HUD, attitude towards the technology markedly improved (M=4.0 vs. M=4.68, p=.002).

Complete data is provided in the Supplemental material.

V. Discussion

We here proposed a methodology to validate the UX in AD systems based on continuous, quantitative information gathered from physiological signals while the user is immersed in a VR driving simulation. Our methodology is exemplified by the comparison of two AR-HUD-based interfaces which differ in the amount of information displayed to the users.

By controlling all aspects of the simulated environment, we were able to disentangle the effect of very specific design choices and measure their impact on the overall UX. It must be stressed that the only difference between the two groups was the information displayed by the HUD, as the simulation was otherwise identical; study groups were also homogeneous in terms of age, sex and ethnicity.

Our results confirmed that providing "why" information is important to reassure the user of the system's competence and to promote trust and situational awareness [3], [9]. To the best of our knowledge, ours is the first contribution to evaluate a realistic HUD displaying a wide range of visual and auditory cues about the vehicle and its surroundings, as it is expected in future AVs. Given the number of objects involved in realistic scenarios, an omni-comprehensive (OMN) display could lead to an excessive cognitive load. A possible way to reduce information load, which we denoted as selective (SEL), is to display only the most relevant visual cues in the current context. Indeed, our results indicated that the users found information displayed by the OMN HUD slightly excessive, although acceptable in both cases, but this was compensated by a less stressful driving experience, as confirmed both by subjective and objective measures.

This difference is especially evident when potentially dangerous events occur, such as a pedestrian crossing the street at the last minute. It is worth noting how the HMI influenced the perception of external events, on one hand, and of the vehicle's performance, on the other hand, despite the fact that the simulated scenario was identical in those respects. For instance, users in the OMN group perceived the vehicle as better equipped to deal with unexpected changes in the environment. We argue that this difference arose as a consequence of the mental model that users formed: as the information provided by the HUD allowed users to better anticipate dangerous situations, they projected this feeling onto the AV as well.

Our results have important implications for AI research in AD, and specifically for the sensory sub-systems, as HMI constraints need to be considered in their design. For instance, end-to-end training from sensory input to planning does not explicitly extract all the information that was included in this simulated HMI [54]. In our simulation, information displayed by the SEL HUD was chosen based on a set of heuristics that could be further improved by exploiting a more advanced AI, such as the ability to predict the motion of objects and pedestrians to foresee potentially dangerous situations before they actually affect the vehicle's trajectory.

In this study, we have sought to be as independent as possible from specific AD systems, e.g., by simulating perfect vehicle sensing capabilities. Our conclusions are thus unaffected by potential errors or misses in the AD object detection system. The proposed methodology could certainly be employed to test other types of autonomous vehicles and their underlying AI systems, by changing the modeled interior and/or behavior. It would also be possible to investigate how possible errors may affect the UX and trust.

The proposed scenario is certainly representative of the learning phase as defined in [9]. Information display by the HUD is particularly relevant in this initial phase, when the user is still forming a mental model of how the AD system works. Our results may not apply entirely to the performance phase, in which the user has observed the AD system for a prolonged period of time. However, the unexpected events or accidents which we simulate, while rare, can have a profound effect on trust, both at the individual and collective level. It should be noticed that trust begins to form even before the first interaction with the system, e.g. based on information from the media, or personal preferences [8], [9]. This was evident in our study where, initially, many subjects were not willing to participate in a real AD experience. However, participating in the VR experience, and being exposed to an informative interface, significantly improved their acceptance towards AD systems. In a simulated setting, all AD technologies, as well as all types of events, can be recreated, opening interesting opportunities for "training" future users of AV technology.

GSR proved capable of detecting user's stress in response to potentially dangerous events, in line with previous literature results which, however, were obtained in the context of manual or partially automated driving [4], [26]. Notably, differences in HMI design were reflected in observable changes in GSR levels, even when using consumer electronics sensors. The GSR response was correlated to the perceived risk as measured by subjective questionnaires, as well as to the "surprise" factor, which depends on the HMI. We here focused on the response to specific events, but the methodology could be extended to extract features that characterize the entire experience [26].

VI. Conclusions and Future Work

In this work we proposed a methodology to validate the UX in AD systems based on continuous, quantitative information gathered from physiological signals while the user is immersed in a VR driving simulation. Its effectiveness was shown in the context of HMI design, and specifically applied to the comparison of HUD-based interfaces for AVs that provides visual cues about the vehicle's sensory and planning systems. We explored in this exemplification the role of HMI in eliciting a sense of trust and safeness in AD systems, as this will be key for humans to relinquish control of the vehicle.

The proposed methodology relies on physiological signals (GSR in this specific embodiment) to provide a continuous, quantitative and objective feedback. This is particularly relevant for simulation of AD systems, as objective measures in driving research are traditionally based on driver's performance and behavior. A limitation of GSR is that it measures arousal, but is a poor indicator of valence. In our specific case, the experience was engineered to elicit a sense of distress and, hence, a positive valence was excluded. In the future, this lack could be overcome by including other sensors, e.g., to measure the HR, other types of features that reflect different characteristics of the UX, as well as machine learning models to more accurately detect the passengers' affective state.

It should be noticed that the increasing adoption of wearable devices like smart watches incorporating a growing set of health sensors will open additional opportunities for AVs' personalization; anthropomorphism, customization and adaptivity are also important factors for trust-worthy HMI [9]. While the physiological response (1–5s in the case of GSR) is too slow to be exploited for actual driving, it could be used to customize various aspects of the HMI, like the quantity and quality of information displayed, and of the overall driving experience.

The proposed methodology for testing could be extended to cover also the above scenarios as well as other aspects of the UX (e.g., considering not just in-vehicle scenarios, but also vehicle-to-pedestrian interactions [55], long-term performance [9], etc.), by adjusting the simulation, the HMI and/or the vehicle's AI as needed.

Acknowledgment

The authors want to thank Dario Doronzo and Antonello Laurino for their contributions on the system implementation. This research was partly supported by the VR@Polito lab.

References

[1] W. Jiadai, L. Jiajia, and K. Nei, "Networking and communications in autonomous driving: A survey," *IEEE Comm. Surveys & Tutor.*, in press.

[2] A. Smith and M. Anderson, "Americans attitudes toward driverless vehicles," 2017. [Online]. Available: http://www.pewinternet.org/2017/10/04/americans-attitudes-toward-driverless-vehicles/

[3] J. Koo, J. Kwac, W. Ju, M. Steinert, L. Leifer, and C. Nass, "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*, vol. 9, no. 4, pp. 269–275, 2015.

[4] L. Eudave and M. Valencia, "Physiological response while driving in an immersive virtual environment," in *Wearable and Implantable Body Sensor Networks (BSN), 2017 IEEE 14th International Conference on*. IEEE, 2017, pp. 145–148.

[5] R. Jose, G. A. Lee, and M. Billinghurst, "A comparative study of simulated augmented reality displays for vehicle navigation," in *Proceedings of the 28th Australian Conference on Computer-Human Interaction*, ser. OzCHI '16, 2016, pp. 40–48.

[6] P. Lungaro, K. Tollmar, and T. Beelen, "Human-to-ai interfaces for enabling future onboard experiences," in *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications Adjunct*, ser. AutomotiveUI '17, 2017, pp. 94–98.

[7] D. Harrison McKnight and N. L. Chervany, "Trust and distrust definitions: One bite at a time," in *Trust in Cyber-societies*, R. Falcone, M. Singh, and Y.-H. Tan, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 27–54.

[8] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Hum. Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[9] F. Ekman, M. Johansson, and J. Sochor, "Creating appropriate trust in automated vehicle systems: A framework for hmi design," *IEEE Trans. Hum-Mach. Syst.*, vol. 48, no. 1, pp. 95–101, 2018.

[10] F. M. F. Verberne, J. Ham, and C. J. H. Midden, "Trust in smart systems: Sharing driving goals and giving information to increase trustworthiness and acceptability of smart systems in cars," *Hum. Factors*, vol. 54, no. 5, pp. 799–810, 2012.

[11] R. Häuslschmid, M. von Bülow, B. Pfleging, and A. Butz, "Supporting trust in autonomous driving," in *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, 2017, pp. 319–329.

[12] A. Doshi, S. Y. Cheng, and M. M. Trivedi, "A novel active heads-up display for driver assistance," *IEEE Trans. Syst., Man, Cybern. B*, vol. 39, no. 1, pp. 85–93, 2009.

[13] Z. Medenica, A. L. Kun, T. Paek, and O. Palinko, "Augmented reality vs. street views: a driving simulator study comparing two emerging navigation aids," in *Proc. 13th Int. Conf. on Human Computer Int. with Mobile Dev. and Serv.* ACM, 2011, pp. 265–274.

[14] S. Kim and A. K. Dey, "Simulated augmented reality windshield display as a cognitive mapping aid for elder driver navigation," in *Proceedings of the SIGCHI Conference on Hum. Factors in Computing Systems*, ser. CHI '09. New York, NY, USA: ACM, 2009, pp. 133–142.

[15] B.-J. Park, J.-W. Lee, C. Yoon, and K.-H. Kim, "Augmented reality for collision warning and path guide in a vehicle," in *Proceedings of the 21st ACM Symposium on Virtual Reality Software and Technology*, ser. VRST '15. New York, NY, USA: ACM, 2015, pp. 195–195.

[16] R. Haeuslschmid, L. Schnurr, J. Wagner, and A. Butz, "Contact-analog warnings on windshield displays promote monitoring the road scene," in *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, ser. AutomotiveUI '15. New York, NY, USA: ACM, 2015, pp. 64–71.

[17] M. T. Phan, I. Thouvenin, and V. Frmont, "Enhancing the driver awareness of pedestrian using augmented reality cues," in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, Nov 2016, pp. 1298–1304.

[18] L. Guo, S. Manglani, Y. Liu, and Y. Jia, "Automatic sensor correction of autonomous vehicles by human-vehicle teaching-and-learning," *IEEE Trans. on Vehicular Technology*, vol. 67, no. 9, pp. 8085–8099, 2018.

[19] F. Bazzano, F. Gentilini, F. Lamberti, A. Sanna, G. Paravati, V. Gatteschi, and M. Gaspardone, "Immersive virtual reality-based simulation to support the design of natural human-robot interfaces for service robotic applications," in *Augmented Reality, Virtual Reality, and Computer Graphics - Third International Conference, AVR 2016, Lecce, Italy, June 15-18, 2016. Proceedings, Part I*, 2016, pp. 33–51.

[20] Y. Chen, C. Stout, A. Joshi, M. L. Kuang, and J. Wang, "Driver-assistance lateral motion control for in-wheel-motor-driven electric ground vehicles subject to small torque variation," *IEEE Trans. on Vehicular Technology*, vol. 67, no. 8, pp. 6838–6850, 2018.

[21] Y. Wang, B. Mehler, B. Reimer, V. Lammers, L. A. D'Ambrosio, and J. F. Coughlin, "The validity of driving simulation for assessing differences between in-vehicle informational interfaces: A comparison with field testing," *Ergonomics*, vol. 53, no. 3, pp. 404–420, 2010.

[22] N. Mullen, J. Charlton, A. Devlin, and M. Bedard, *Simulator validity: behaviours observed on the simulator and on the road*, 1st ed. Australia: CRC Press, 2011, pp. 1 – 18.

[23] S. Balters and M. Steinert, "Capturing emotion reactivity through physiology measurement as a foundation for affective engineering in engineering design science and engineering practices," *J. of Intell. Manuf.*, vol. 28, no. 7, pp. 1585–1607, 2017.

[24] D. Ruscio, L. Bascetta, A. Gabrielli, M. Matteucci, D. Ariansyah, M. Bordegoni *et al.*, "Collection and comparison of driver/passenger physiologic and behavioural data in simulation and on-road driving," in *Models and Technologies for Intelligent Transportation Systems (MT-ITS), 2017 5th IEEE International Conference on*, 2017, pp. 403–408.

[25] M. J. Johnson, T. Chahal, A. Stinchcombe, N. Mullen, B. Weaver, and M. Bedard, "Physiological responses to simulated and on-road driving," *Int. J. Psychophysiol.*, vol. 81, no. 3, pp. 203–208, 2011.
[26] J. Healey, R. W. Picard *et al.*, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 2, pp. 156–166, 2005.
[27] R. R. Singh, S. Conjeti, and R. Banerjee, "Assessment of driver stress from physiological signals collected under real-time semi-urban driving scenarios," *Int. J. of Comput. Int. Sys.*, vol. 7, no. 5, pp. 909–923, 2014.
[28] B. Dalgarno and M. J. Lee, "What are the learning affordances of 3-d virtual environments?" *Brit. J. Ed. Tech.*, vol. 41, no. 1, pp. 10–32, 2010.
[29] Genivi Vehicle Simulator. [Online]. Available: https://at.projects.genivi.org/wiki/display/PROJ/GENIVI+Vehicle+Simulator
[30] R. Marino, S. Scalzi, and M. Netto, "Nested pid steering control for lane keeping in autonomous vehicles," *Control Engineering Practice*, vol. 19, no. 12, pp. 1459–1467, 2011.
[31] A. Laurino, "Virtual reality-based simulation tools for evaluating user experience in autonomous vehicles," Master's thesis, 2018.
[32] P. A. Hancock, R. J. Jagacinski, R. Parasuraman, C. D. Wickens, G. F. Wilson, and D. B. Kaber, "Human-automation interaction research: Past, present, and future," *Ergon. in Design*, vol. 21, no. 2, pp. 9–14, 2013.
[33] S. W. A. Dekker and D. D. Woods, "Maba-maba or abracadabra? progress on human-automation co-ordination," *Cogn. Technol. Work*, vol. 4, pp. 240–244, 2002.
[34] ISO, "Intelligent transport systems – forward vehicle collision warning systems – performance requirements and test procedures," ISO 22324:2015, 2013.
[35] A. Sebastian, M. Tang, Y. Feng, and M. Looi, "Multi-vehicles interaction graph model for cooperative collision warning system," in *2009 IEEE Intelligent Vehicles Symposium*, June 2009, pp. 929–934.
[36] G. Johansson and K. Rumar, "Drivers' brake reaction times," *Hum. Factors*, vol. 13, no. 1, pp. 23–27, 1971.
[37] P. George, I. Thouvenin, V. Frmont, and V. Cherfaoui, "Daaria: Driver assistance by augmented reality for intelligent automobile," in *2012 IEEE Intelligent Vehicles Symposium*, June 2012, pp. 1043–1048.
[38] BMW Group, Jun 2018. [Online]. Available: https://www.press.bmwgroup.com/global/article/detail/T0197906EN/pedestrian-warning-with-city-braking-activation-from-bmw-connected\drive-receives-euro-ncap-advanced-award-2014?language=en
[39] Volvo Cars, May 2018. [Online]. Available: https://www.media.volvocars.com/uk/en-gb/media/videos/18528
[40] M. Ferati, P. Murano, and G. A. Giannoumis, "Universal design of user interfaces in self-driving cars," in *International Conference on Applied Hum. Factors and Ergonomics*, 2017, pp. 220–228.
[41] B. Kim, D. Kim, K. Kim, and K. Yi, "High-level automated driving on complex urban roads with enhanced environment representation," in *2015 15th International Conference on Control, Automation and Systems (ICCAS)*, Oct 2015, pp. 516–521.
[42] G. Valenza, A. Lanata, and E. P. Scilingo, "The role of nonlinear dyn. in affective valence and arousal recognition," *IEEE transactions on affective computing*, vol. 3, no. 2, pp. 237–249, 2012.
[43] M. Meehan, B. Insko, M. Whitton, and F. P. Brooks Jr, "Physiological measures of presence in stressful virtual environments," in *ACM Trans. Graphics (tog)*, vol. 21, no. 3, 2002, pp. 645–652.
[44] B. Figner, R. O. Murphy *et al.*, "Using skin conductance in judgment and decision making research," *A handbook of process tracing methods for decision research*, pp. 163–184, 2011.
[45] M. Slater, C. Guger, G. Edlinger, R. Leeb, G. Pfurtscheller, A. Antley *et al.*, "Analysis of physiological responses to a social situation in an immersive virtual environment," *Presence: Teleoperators and Virtual Environments*, vol. 15, no. 5, pp. 553–569, 2006.
[46] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator sickness questionnaire: an enhanced method for quantifying simulator sickness," *Int. J. Aviat. Psychol.*, vol. 3, pp. 203–220, 1993.
[47] R. Taylor, "Situational awareness rating technique: The development of a tool for aircrew systems design," in *Sit. Awar.*, 2017, pp. 111–128.
[48] NASA. [Online]. Available: https://humansystems.arc.nasa.gov/groups/TLX/
[49] K. E. Schaefer, "Measuring trust in human robot interactions: Development of the trust perception scale-hri," in *Robust Intelligence and Trust in Autonomous Systems*. Springer, 2016, pp. 191–218.
[50] R. S. Kalawsky, "VRUSE - a computerised diagnostic tool: for usability evaluation of virtual/synthetic environment systems," *Appl. Ergon.*, vol. 30, no. 1, pp. 11–25, 1999.
[51] "Grove GSR sensor Seeed Wiki," 2017. [Online]. Available: http://wiki.seeedstudio.com/Grove-GSR_Sensor/
[52] "Raspberry ADC," 2015. [Online]. Available: https://learn.adafruit.com/raspberry-pi-analog-to-digital-converters/mcp3008
[53] B. Patrao, S. Pedro, and P. Menezes, "How to deal with motion sickness in virtual reality," *Sciences and Tech. of Int., 2015 22nd*, pp. 40–46, 2015.
[54] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, *et al.*, "End to end learning for self-driving cars," *arXiv:1604.07316*, 2016.
[55] R. Amir and K. T. John, "Autonomous vehicles that interact with pedestrians: A survey of theory and practice," *IEEE Transactions on Intelligent Transportation Systems*, in press.

Lia Morra received the M.Sc. and the Ph.D. degrees in computer engineering from Politecnico di Torino, Italy, in 2002 and 2006. Currently, she is senior post-doctoral fellow at the Dip. di Automatica e Informatica of Politecnico di Torino. Her research interests include computer vision, pattern recognition, and machine learning.

Fabrizio Lamberti is an associate professor at the Dip. di Automatica e Informatica of Politecnico di Torino. His research interests are manly in the areas of computer graphics, HMI and intelligent computing. He is serving as Associate Editor for the IEEE Transactions on Computers, the IEEE Transactions on Emerging Topics in Computing, the IEEE Transactions on Learning Technologies and the IEEE Transactions on Consumer Electronics. He is a Senior Member of the IEEE.

F. Gabriele Pratticó received his M.Sc. degrees in computer engineering from Politecnico di Torino, Italy, in 2017. Currently, he is a Ph.D. student at Politecnico di Torino, where he carries out research in the areas of mixed reality, HMI, serious games and user experience design.

Salvatore La Rosa received the M.Sc. degree in biomedical engineering from Politecnico di Torino, Italy, in 2019. His major interests regard biosignal analysis, pattern recognition, machine learning and embedded systems.

Paolo Montuschi is a full professor at the Dip. di Automatica e Informatica and a Member of the Board of Governors of Politecnico di Torino, Italy. His research interests include computer arithmetic, computer graphics, and intelligent systems. He is serving as 2019 Acting (interim) Editor-in-Chief of the IEEE Transactions on Emerging Topics in Computing and as the 2017-19 IEEE Computer Society Awards Chair. He is an IEEE Fellow, and a life member of the International Academy of Sciences of Turin and of IEEE Eta Kappa Nu.