Context-Adaptive Management of Drivers' Trust in Automated Vehicles

Hebert Azevedo-Sa, Suresh Kumaar Jayaraman, X. Jessie Yang, Lionel P. Robert Jr., and Dawn M. Tilbury

***Abstract*—Automated vehicles (AVs) that intelligently interact with drivers must build a trustworthy relationship with them. A calibrated level of trust is fundamental for the AV and the driver to collaborate as a team. Techniques that allow AVs to perceive drivers' trust from drivers' behaviors and react accordingly are, therefore, needed for context-aware systems designed to avoid trust miscalibrations. This letter proposes a framework for the management of drivers' trust in AVs. The framework is based on the identification of trust miscalibrations (when drivers' undertrust or overtrust the AV) and on the activation of different communication styles to encourage or warn the driver when deemed necessary. Our results show that the management framework is effective, increasing (decreasing) trust of undertrusting (overtrusting) drivers, and reducing the average trust miscalibration time periods by approximately 40%. The framework is applicable for the design of SAE Level 3 automated driving systems and has the potential to improve the performance and safety of driver–AV teams.**

***Index Terms*—Intelligent transportation systems, social human-robot interaction, human factors and human-in-the-loop.**

I. Introduction

TRUST influences the interactions between people and automated systems [1]. In this letter, trust is defined as "*the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability*" [2]. In the future, automated systems will be expected to become aware of humans' trusting behaviors and to adapt their own behaviors, seeking to improve their interaction with humans [3]. One way to implement those adaptive capabilities is to develop methods for trust *management*, which we consider to be a robot's ability to estimate and, if needed, to recalibrate a human's trust in that robot. Trust miscalibration is defined as a mismatch between a human's trust in an automated system and the capabilities of that system [4], [5]. Trust miscalibration is characterized by *overtrusting* or *undertrusting* an automated system, and it can harm the performances associated with the use of that system. Overtrusting an automated system can lead to *misuse*, where the human user relies on the system to handle tasks that exceed its capabilities. Undertrusting an automated system can lead to *disuse*, where the human fails to fully leverage the system's capabilities. Proper trust management can avoid both misuse and disuse of the automated system by estimating and, if needed, influencing the human's trust in the system to avoid trust miscalibration.

The ability to manage trust and avoid miscalibration is especially crucial for automated systems that can put people's lives at risk, such as automated vehicles (AVs). Either misuse or disuse of an AV is a risk to the performance and safety of the team formed by the driver and the AV. Considering the current technology race in the automotive industry for AV development [6], AVs that can manage drivers' trust are a significant—if not urgent—demand. In the driver–AV interaction context, the trust management problem consists of two main challenges: 1) the accurate estimation of trust in the AV and 2) the recalibration or realigning of the driver's trust with the AV's capabilities. The goal of trust estimation is to provide accurate real-time estimates of the drivers' trust in the AV based on behavioral cues. The goal of trust calibration is to set the driver's trust to appropriate levels through a trust influence mechanism, for instance, by adapting the communication between the AV and the driver.

Previous research has obtained promising results regarding trust estimation in AVs using drivers' behaviors and actions [7]–[9]. The second challenge, however, has not received as much attention. In fact, we know of no prior research that has addressed the problem of trust calibration, with the goal of manipulating trust in AVs to avoid undertrust or overtrust in real time. This letter presents a framework for managing trust in AVs, focusing on how to recalibrate drivers' trust after a trust miscalibration has been identified. Our framework integrates a Kalman filter-based trust estimator developed in previous work [9] and an unprecedented real-time *trust calibrator*. We draw inspiration from recent approaches that have provided valuable insights for the development of trust estimators [7], [8]. These approaches, however, fall short on presenting strategies for adapting the behavior of the AV and manipulating the driver's trust to, ultimately, improve the driver–AV team performance through trust calibration.

Manuscript received May 6, 2020; accepted September 1, 2020. Date of publication September 22, 2020; date of current version September 30, 2020. This letter was recommended for publication by Associate Editor H. Myung and Editor Y. Choi upon evaluation of the Reviewers' comments. This work was supported in part by the National Science Foundation, in part by the Brazilian Army's Department of Science and Technology, and in part by the Automotive Research Center (ARC) at the University of Michigan, with funding from government contract DoD-DoA W56HZV14-2-0001, through the U.S. Army Combat Capabilities Development Command (CCDC)/Ground Vehicle Systems Center (GVSC). *(Corresponding author: Hebert Azevedo-Sa.)*

Hebert Azevedo-Sa, X. Jessie Yang, and Lionel P. Robert Jr. are with the Robotics Institute, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: azevedo@umich.edu; xijyang@umich.edu; lprobert@umich.edu).

Suresh Kumaar Jayaraman and Dawn M. Tilbury are with the Mechanical Engineering Department, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: jskumaar@umich.edu; tilbury@umich.edu).

This article has supplementary downloadable material available at https://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/LRA.2020.3025736

Fig. 1. An undertrusting driver is encouraged by the AV system simulator to focus on his non-driving-related task (NDRT), to increase his trust level. An analogous situation would take place if the driver overtrusted the AV's capabilities, with the system then demanding his attention to the driving task.

Our contribution is a rule-based trust calibrator that can be integrated with previously proposed trust estimators. With this integration, we introduce the novelty of a trust management framework. The trust calibrator compares the AV's capabilities with the driver's trust estimates to identify trust miscalibrations, and modifies the interactive behavior of the AV accordingly. The AV is the element that directly interacts with the drivers, providing verbal messages intended to influence drivers' trust in the AV. We validated our trust management framework on a user study with 40 participants, where the participants operated an AV simulator while simultaneously performing a non-driving-related task (NDRT) and having their behavior observed (from which their trust levels were estimated). Our results show that the proposed framework was successful in its intent, being able to increase trust levels when drivers undertrusted the system and to decrease trust levels when drivers overtrusted the system. With the proposed trust calibrator, our management framework reduced the time periods for which trust was miscalibrated by approximately 40%. Consequently, the method introduced in this letter mitigates the occurrence of unsafe driving scenarios and generally improves the driver–AV team performance and safety. Fig. 1 illustrates our study situation, where an undertrusting driver is exhorted by the AV system to focus on his NDRT (with the objective to increase his trust level). An analogous situation would take place if the driver overtrusted the AV's capabilities, with the system demanding his attention to the driving task.

This letter's remaining content is as follows. Section II provides the theoretical bases for trust management in the driver–AV interaction context. Section III presents our experimental methodology. In Section IV, we discuss the results of the study. Finally, Section V concludes the letter and presents our suggestions for future work.

II. Management of Trust in AVs

A. Related Work: Trust in Automated Systems, Trust Estimation and Trust Calibration

Trust has been long discussed as a factor that mediates the interaction between humans and automated systems in the field of supervisory control [4], [5], [10]–[13]. Researchers have established formal definitions that evolved from social science's original descriptions of trust in interpersonal relationships [14], [15]. Measuring trust in automated systems is a challenging task because trust is an abstract concept that depends both on the context and on the individual *trustor*. That challenge led to the establishment of standard scales for measuring trust [13], [16]. When using these scales, the measurement procedure relies on users' self-reports, which have clear practical limitations when researchers are interested in tracking trust levels for real-time applications. Given these limitations, techniques for trust estimation that can take advantage of models for trust dynamics have been investigated [7], [17]–[21].

Trust estimation is the first challenge to be overcome in a trust management framework. To avoid collecting self-reports from users, systems have to use advanced perception techniques to process users' behaviors and actions. For instance, eye–tracking has been used for estimating trust in unmanned aerial vehicle controllers [22]. Specifically for AVs, researchers have worked with physiological signals (i.e., electroencephalography and galvanic skin response) to develop a classifier-based empirical trust sensor [7]. The privileged sensing framework (PSF) was applied with that same type of physiological signals to anticipate and influence humans' behaviors, with the goal of optimizing changes in control authority between the human and the automated system [8], [23]. Classic methods, such as Kalman filtering, have also been used for trust estimation [9].

Trust calibration is as important as trust estimation and plays a fundamental role in trust management. In this study, the objective of trust calibration was to manipulate drivers' trust in the AV for aligning trust with the AV's capabilities (i.e., avoiding trust miscalibration). Several studies have identified factors that significantly impact trust in AVs, and, therefore, could be used for trust manipulation purposes. The most important of these factors are situation awareness and risk perception, which are influenced by the ability of the AV to interact with the driver. For instance, enhancing drivers' situation awareness facilitates increased trust in AVs [24], [25]. On the other hand, increasing drivers' perception of risk reduces their trust in AVs [26]–[28]. Our framework takes advantage of these studies' results and seeks to influence trust by varying situation awareness and risk perception through verbal communications from the AV to the driver.

B. Problem Statement

Considering the context of a driver interacting with an AV featuring an SAE Level 3 automated driving system (ADS), we addressed two main problems. First, we aimed to identify instances for which drivers' trust in the AV is miscalibrated, i.e., when the driver is undertrusting or overtrusting the AV. Second, we focused on manipulating drivers' trust in the AV to achieve calibrated levels, i.e., trust levels that match the AV's capabilities [2]. In other words, our goal was to increase or decrease drivers' trust in the AV whenever drivers were undertrusting or overtrusting the AV, respectively.

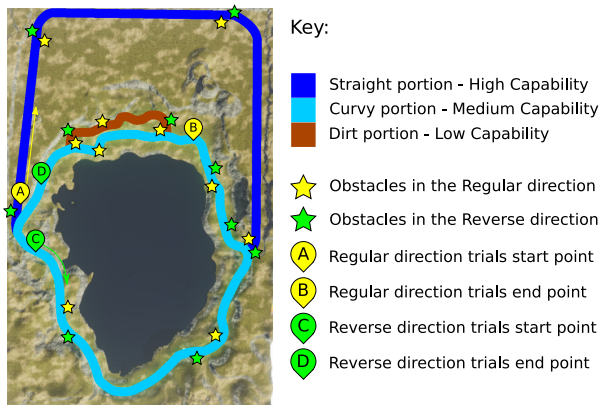

Fig. 2. Circuit track used in this study. The portions of the road correspond to the capability of the AV. In the regular direction, drivers start at point A, follow the "straight" path in the clockwise direction, cover the curvy path and finish the trial at point B, right after passing through the dirt road portion. In the reverse direction, drivers start at point C, follow the curvy path in counterclockwise direction, cover the straight path, continue to the curvy path (until the dirt portion), pass through the dirt portion, and finish the trial at point D. Both directions have 12 events (encounters with obstacles), and it took drivers approximately 10 to 12 minutes to complete a trial.

In SAE Level 3 ADSs, drivers are required to take back control when the system requests intervention or when it fails [29]. We assume that the AV features automated lane keeping, cruise control and forward collision alarm functions that can be activated (all at once) and deactivated at any time by the driver. The AV can also identify different road difficulty levels and process drivers' behavioral signals to estimate their trust in the AV.

C. Solution Approach

We implemented a scenario to represent the problem context described in Subsection II-B with an AV simulator. We established simulations where drivers took trials in a predefined circuit track. The circuit track was divided into distinct parts, having three predefined risk levels, corresponding to the difficulty associated with each part of the circuit track. The easy parts of the circuit track consisted of predominantly straight roads; the intermediate difficulty parts were curvy paved roads; and the difficult parts were curvy dirt roads. Within these trials, drivers encountered abandoned vehicles on the road, which represented obstacles that the AV was not able to maneuver around by itself (using its automated driving functions). At that point, drivers had to take over control, pass the obstacle and then engage the autonomous driving mode again. Fig. 2 shows the circuit implemented in the simulation environment.

We needed to compare drivers' trust levels and the AV's capability levels to identify trust miscalibrations. Therefore, we defined three capability levels for the AV, corresponding to the difficulty of the circuit track parts. The AV's forward collision alarm was able to identify the obstacles and also to trigger an emergency brake if the driver did not take control in time to maneuver around the obstacles. These two actions were activated at different distances to the obstacles, represented by the two circular regions represented in Fig. 3. On straight paved roads these distances were larger, representing the longer perception

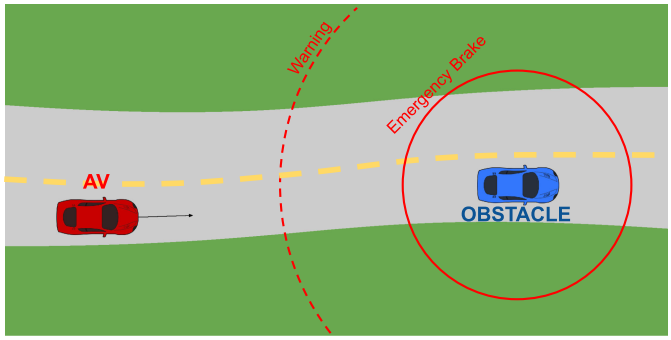

Fig. 3. Concentric circles represent the distances for which the warning message "Stopped vehicle ahead!" was provided to the driver, and the emergency brake was triggered. The distances varied according to the difficulty of the road. If the emergency brake was triggered, the drivers were penalized on their NDRT score.

ranges of the AV sensors. On more difficult parts of the circuit (i.e., curvy or dirt), however, the curves and the irregular terrain reduced that perception range, implying shorter distances. The AV was able to identify the obstacle, warn the driver and eventually brake at a fair distance from the obstacle when it was operated on straight roads. This condition corresponded to the AV's high capability. On curvy and dirt roads, the AV was not able to anticipate the obstacles at a reasonable distance, giving drivers less time to react and avoid triggering the emergency brake. These conditions corresponded to the AV's medium and low capabilities.

In the scenario, drivers also had to simultaneously perform a visually demanding NDRT, consisting of a visual search on a separate touchscreen device that exchanged information with the AV. They performed the NDRT only when the self-driving capabilities were engaged. The behavioral measures taken from the drivers were their focus on the NDRT (from an eye tracker); their ADS usage rate; and their NDRT performance, measured by the number of correctly performed visual searches per second. Drivers were penalized if the emergency brake was triggered, which gave them a sense of costs and risks of neglecting the AV operation. Specific details about the tasks are given in Section III-C.

The block diagram in Fig. 4 presents our proposed trust management framework, composed of two main blocks: the trust estimator and the trust calibrator. The AV block represents elements of the vehicle, such as the sensors to monitor the environment and the ability to output verbal messages to interact with the driver. We present the definitions and the notation used in this letter in Table I.

D. Trust Estimator

Fig. 4 illustrates the trust estimator block, with the AV's alarms and the observation variables φ_k, υ_k and π_k as inputs, and a numerical estimate of drivers' trust in the AV as the output T_k. The observation variables capture the drivers' behavior, which is affected by drivers' trust in the AV. This trust estimator is a simplified version of what is presented in [9], and was chosen because of its simple implementation and proven ability to track

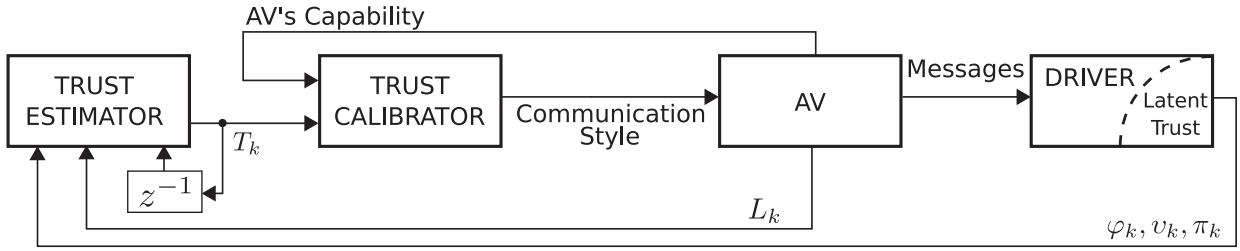

Fig. 4. Block diagram that represents the trust management framework. The trust estimator block provides a trust estimate T_k to the trust calibrator, which compares it to the capabilities of the AV during operation. The calibrator then defines the communication style that the AV should adopt, and the AV provides the corresponding verbal messages to the driver. L_k represents an alarm provided by the ADS when an obstacle on the road is identified. The observation variables φ_k, υ_k and π_k represent drivers' behaviors, from which drivers' "real" trust (considered a latent variable) is estimated. A delay of one event is represented by the z^{-1} block.

TABLE I
DEFINITIONS AND NOTATION

Definition, notation	Characterization
Trial, $[t_0, t_f] \in \mathbb{R}^+$	Trials occur when drivers operate the vehicle on a predefined route, and are characterized by their corresponding time intervals.
Events, $k \in \mathbb{N} \setminus \{0\}$	Events occur each time the ADS warns the driver about an obstacle on the road at t_k, $t_0 < t_k < t_f$.
Alarm, $L_k \in \{0, 1\}$	Boolean variable that is set when the AV correctly identifies an obstacle and warns the driver at the event k. It is reset after the driver passes the obstacle
Focus, $\varphi_k \in [0, 1]$	Drivers' focus on the NDRT, the ratio of time the driver spends looking at the NDRT screen during $[t_k, t_{k+1})$.
Usage, $\upsilon_k \in [0, 1]$	Drivers' ADS usage, the ratio of time the driver spends using the AV's self-driving capabilities during $[t_k, t_{k+1})$.
Performance, $\pi_k \in \mathbb{R}$	Drivers' NDRT performance, the number of points obtained on the NDRT during $[t_k, t_{k+1})$, divided by $\Delta t_k = t_{k+1} - t_k$.
Trust in the AV, $T_k \in [0, 100]$	Drivers' estimated trust in the AV. It is assigned to the interval $[t_k, t_{k+1})$, computed from φ_k, υ_k, π_k and is associated with the covariance Σ_T.

drivers' trust. Alternative trust estimators could be integrated to the proposed trust management framework if the inputs they require can be captured in real-time. Differently from [9], we considered that the alarms L_k were always reliable (true alarms), and could not be false alarms or misses. For the sake of completeness, we briefly describe the trust dynamics model used in this study.

The discrete LTI state-space model for trust dynamics has the form (1),

$$T_{k+1} = \mathbf{A}T_k + \mathbf{B}L_k + u_k, \tag{1a}$$

$$\begin{bmatrix} \varphi_k \\ \upsilon_k \\ \pi_k \end{bmatrix} = \mathbf{C}T_k + \mathbf{w}_k. \tag{1b}$$

T_{k+1}, the trust estimate at the event $k+1$, depends on T_k, the alarm L_k, and the process noise u_k. The observation variables depend on the estimated trust and output noise $\mathbf{w}_k$. $\mathbf{A} = \begin{bmatrix} 1.0 \end{bmatrix}$; $\mathbf{B} = \begin{bmatrix} 0.40 \end{bmatrix}$; $\mathbf{C} = 10^{-3} \times \begin{bmatrix} 7.0 & 4.2 & 9.2 \end{bmatrix}^\top$; $u_k \sim \mathcal{N}(0, 0.25^2)$; and $\mathbf{w}_k \sim \mathcal{N}(\mathbf{0}, \mathrm{diag}(\sigma_\varphi^2, \sigma_\upsilon^2, \sigma_\pi^2))$, with $\sigma_\varphi = 1.8 \times 10^{-4}$, $\sigma_\upsilon = 7.0 \times 10^{-5}$ and $\sigma_\pi = 5.7 \times 10^{-2}$. (Please see Table I for variables' definitions.) The parameters for (1) are found by fitting linear models [30] using a previously obtained data set. The state-space structure permits the application of Kalman filter-based techniques for the estimator design. The trust estimator is initialized with

$$T_0 = \frac{1}{3}\left(\frac{\varphi_0}{c_1} + \frac{\upsilon_0}{c_2} + \frac{\pi_0}{c_3}\right), \tag{2}$$

where φ_0, υ_0 and π_0 measured over the interval $[t_0, t_1)$ and c_1, c_2, c_3 are the entries of $\mathbf{C}$.

E. Trust Calibration

The trust calibrator block represented in Fig. 4 was intended to affect drivers' situation awareness (or risk perception) by changing the communication style of the AV, with the goal of influencing drivers' trust in the AV [31]. At every event k, the AV interacted with the driver through verbal messages corresponding to the communication style defined in the trust calibrator block. The AV can encourage the driver to focus on the NDRT, moderately warn the driver about the difficulties of the road ahead, or harshly warn the driver, literally demanding driver's attention. Table II presents the messages the AV provided to the driver in four different communication styles.

To identify trust miscalibrations, the trust calibrator compares the trust estimates with the capability of the AV. Lee and See [2] considered both trust in the automated system and the capabilities of the system as continua that must be comparable within each other. We assumed that the AV's capability corresponds to the three difficulty levels of the road where the AV is operated. We divided the interval $[0, 100]$, for which drivers' trust in the AV was defined, into three sub-intervals: $[0, 25)$ corresponding to low trust, $[25, 75)$ corresponding to medium trust and $[75, 100]$ corresponding to high trust. The uneven distribution of the sub-interval lengths was chosen to mitigate the uncertainty involved in trust estimation. We fit a wider range of values in the medium level, and considered as "low trust" or "high trust"

TABLE II
MESSAGES PROVIDED BY THE AV IN EACH COMMUNICATION STYLE

AV Communication Style	Message
Encouraging	"Hey, this is an easy road. You don't need to worry about driving. I will take care of it while you focus on finding the Qs."
Silent	[No message]
Warning (moderate)	"Hey, this part of the road is not very easy. You can still find the Qs, but please pay more attention to the road."
Warning (harsh)	"Look, I told you! I do need your attention. I can feel the road is terrible. I don't know if I can keep us totally safe!"

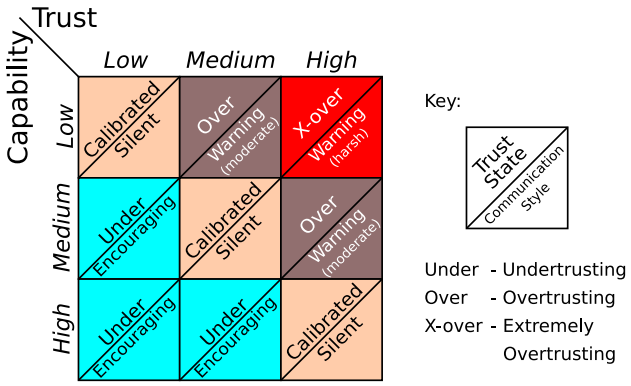

Fig. 5. Rule set for the trust calibrator. The driver's trust state and the communication style are defined when the AV compares its capability and the driver's trust level. E.g.: when trust is lower than the AV's capabilities (light blue cells), the driver is undertrusting the AV, and the encouraging communication style is selected.

only the estimates that were closer to 0 or 100, respectively. The quantization of both the driver's trust in the AV and the AV's capability in three levels facilitates the real-time comparison of these metrics. Moreover, it permits the definition of a finite set of rules for the trust miscalibration issues. Depending on the application context, alternative quantizations or AV capabilities distributions can be implemented without significant changes to the trust calibrator's framework.

A trust miscalibration is identified whenever there is a mismatch between the AV's capabilities and the driver's level of trust in the AV. The communication style of the AV is then selected after the trust miscalibration is identified. At every event, this comparison results in the identification of one of four distinct driver trust states: undertrusting the AV (*Under*); having an appropriate level of trust in the AV (*Calibrated*); overtrusting the AV (*Over*); or extremely overtrusting the AV (*X-over*). Fig. 5 shows the rule set and the correspondence with the resultant communication styles of the AV. Note that the establishment of three levels for trust and AV capability is able to cover the occurrence of both undertrust and overtrust, and also allows the identification of extreme overtrust. Extreme overtrust occurs when a driver has a high level of trust in the AV while the AV's capability is low, which is likely to be crucial for driver safety. Therefore, we consider extreme overtrust a trust miscalibration issue that should be seriously addressed.

III. METHODS

A total of 40 participants ($\mu_{AGE} = 31$; $\sigma_{AGE} = 14$ years) were recruited to take part in the study. From these, 18 were female, 21 male and 1 preferred not to specify gender. We used emails and specialized advertising on a web portal for behavioral and health studies recruitment. All regulatory ethical concerns were taken, and the study was approved by the University of Michigan's Institutional Review Board.

A. Procedure

Participants signed a consent form and filled out a pre-experiment survey as soon as they arrived at the experiment location. Next, the functions of the AV and the experiment dynamics were explained, and a training drive allowed participants to get familiar with both the AV simulator controls and the NDRT. Participants put the eye-tracker device on and, after it was calibrated, started their first trial on the AV simulator. After the trial, they filled out a post-trial survey. Next, they had their second trial and filled out the post trial survey for the second time. Each experiment took approximately 1 h, and the participants were compensated for taking part in the study. The compensation varied accordingly to their highest total number of points obtained in the NDRT, considering both of their trials. Minimum compensation was of \$15, and the participants were able to achieve \$20, \$30 or \$50 in total with a performance cash bonus.

B. Conditions Randomization

All participants experienced one trial with the trust calibrator and one trial without the trust calibrator. To avoid the participants driving in exactly the same conditions in both of their trials, we varied the direction of the driving on the circuit track. Participants drove in clockwise direction (i.e., regular direction) and counterclockwise direction (i.e., reverse direction), as mentioned in Fig. 2. The "trust calibrator use" $\times$ "drive direction" conditions were randomly assigned, depending on the participant's sequential identification number.

C. Tasks and Apparatus

The driving task was implemented with AirSim over Unreal Engine [32]. The visual search NDRT consisted of finding "Q" characters among a field of "O" characters. Participants' score increased by 1 point every time they correctly selected the targets on the screen, and they lost 20 points each time the emergency brake was activated. The NDRT was implemented with PEBL [33]. Source codes for both tasks are available at https://github.com/hazevedosa/tiavManager.

The experimental setup is shown in Fig. 1. The simulator was composed of three screens integrated with a *Logitech* G-27

TABLE III
COMMUNICATION STYLE FIXED EFFECTS ON DRIVERS' TRUST IN AV DIFFERENCE (ΔT), OBTAINED WITH A LINEAR MIXED-EFFECTS MODEL [30]

Trust State / Communication Style	Parameter	Estimate	Standard Error (S.E.)	Student's t	p-value	Lower Bound	Upper Bound
Calibrated / Silent*	β_0	1.7	1.8	0.92	0.36	-1.9	5.3
Under / Encouraging	β_1	**+15.4**	3.3	4.7	3.3×10^{-6}	9.0	21.8
Over / Warning (moderate)	β_2	**-9.0**	2.8	-3.2	1.7×10^{-3}	-14.6	-3.4
X-over / Warning (harsh)	β_3	**-22.9**	5.1	-4.5	9.9×10^{-6}	-33.0	-12.8

Obs.: *Model intercept reference; Significant parameter estimates ($p < 0.01$) in bold font. A random intercept is assigned to each participant in the data set.

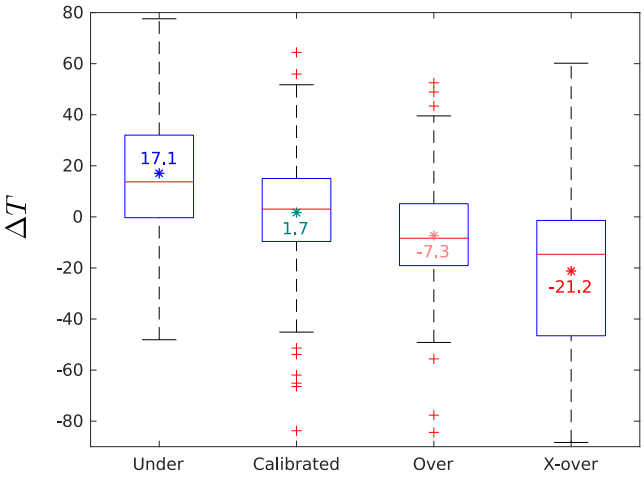

Fig. 6. Distributions of drivers' trust in the AV differences (ΔT), for the different driver trust states. Overtrusting drivers received the warning communication styles and responded with negative differences. Undertrusting drivers received the encouraging communication styles and responded with positive differences. Drivers with calibrated trust had relatively small positive differences on average. The average values were obtained from the parameter estimates in Table III.

driving console, another touch screen for the NDRT and a *Pupil Lab's Mobile* eye-tracker headset.

IV. RESULTS AND DISCUSSION

We analyzed the impacts of using the trust calibrator's adaptive communication with different communication styles on drivers' trust in the AV (i.e., real-time estimated trust T_k). For this, we analyzed the differences in drivers' trust estimates between consecutive events, after they had heard the messages from the AV. Drivers' trust differences are given by $\Delta T = T_k - T_{k-1}$, i.e., the difference between trust estimates after and before the event k. ΔT was specifically computed for the analysis, and indicates how participants' trust estimates changed after they were encouraged or warned by the AV at the event k (i.e., after the AV interacted with the drivers adopting the communication style corresponding to drivers' trust states at the event k).

Drivers showed significant positive or negative differences in their trust estimates after the AV encouraged or warned them. Table III and Fig. 6 present the results obtained with a linear mixed-effects model for ΔT. Linear mixed-effects models are regression models that include both fixed and random effects of independent variables on a dependent variable. Fixed effects represent the influence of the independent variables or treatments of primary interest (in this case, the communication styles) on the dependent variable (i.e., trust difference ΔT). Random effects represent differences that are not explained by the factors of primary interest but are rather related to hierarchical organizations present in the sample population (e.g., groups of data collected from the same participant) [30]. For instance, in this analysis, a random intercept for each participant in the experiment was added to the ΔT linear mixed-effects model. In summary, we sought the β parameters that best fit the model

$$\Delta T = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon_p, \tag{3}$$

where $x_1 = 1$ when the communication style was "Encouraging" and $x_1 = 0$ otherwise; $x_2 = 1$ when the communication style was "Warning (moderate)" and $x_2 = 0$ otherwise; and $x_3 = 1$ when the communication style was "Warning (harsh)" and $x_3 = 0$ otherwise. The random effect ϵ_p had mean $\mu = 0$ and standard deviation $\sigma = 25.3$, and represented each participant's characteristic intercept and the irreducible error of the model. Table III shows that all β parameter estimates corresponding to the non-silent communication styles were significant ($p < 0.01$).

In general, the reaction of the drivers to the AV messages followed an expected trend. The lack of messages did not significantly change driver's behaviors when their trust in the AV was calibrated: the average difference—considered the reference intercept for the linear mixed-effects model—was 1.7 units, but the p-value of 0.36 indicates that it was not significantly different from 0. The encouraging messages helped drivers to increase their trust in the AV: as shown in Table III, the average increase was $1.7 + 15.4 = 17.1$ units for undertrusting drivers. The warning messages had the effect of decreasing their trust in the AV: trust estimates of overtrusting drivers varied by $1.7 - 9.0 = -7.3$ units, and for extremely overtrusting drivers, trust estimates varied by $1.7 - 22.9 = -21.2$ units. Fig. 7 exemplifies the time trace for a participant's trust estimates during a trial, indicating the messages provided by the AV and the regions for which trust would be considered calibrated.

The use of the calibrator reduced trust miscalibrations for 29 (out of 40) participants. We computed *trust miscalibration time ratios*, representing the amount of time drivers' trust state was different from "Calibrated," relative to the total time of each trial. For the computation, we removed the intervals right after a change in AV's capabilities, where miscalibrations were intentionally caused. For all participants, the average trust miscalibration time ratio was 70% in trials for which the calibrator was not used. This ratio was reduced to 43.7% when the calibrator was used. Considering only the 29 participants that

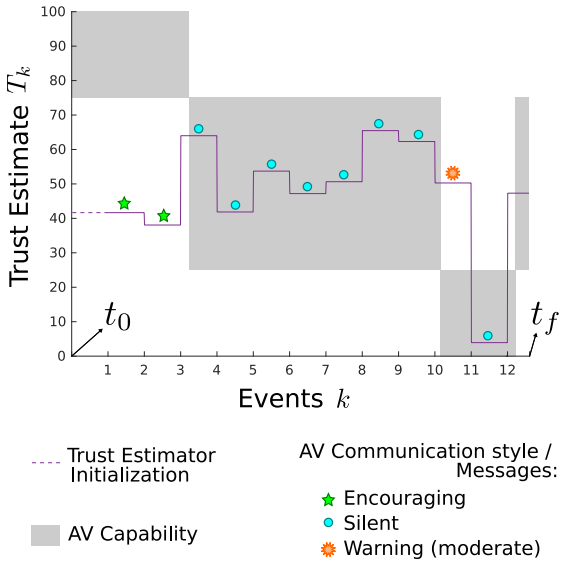

Fig. 7. Time trace for a driver's trust estimates T_k, which is assigned to the interval $[t_k, t_{k+1})$ after being computed from φ_k, υ_k and π_k. After two encouraging messages when the driver undertrusted the AV, T_k increased. After a warning message when the driver overtrusted the AV, T_k decreased. While driver's trust was calibrated, the calibrator refrained from providing messages to the driver.

had their miscalibration time ratios reduced (when using the trust calibrator), these ratios were 82% and 42%, respectively. For the remaining 11 participants, the reasons for their lack of decreased trust miscalibration are unknown, although we believe these reasons could be related to the limitations imposed by the short duration of the experiment.

These results support the effectiveness of our trust management framework or, more specifically, our trust calibrator, which is the main intended contribution of this letter. When undertrusting drivers increase their trust in the AV, their trust state is likely to approach the condition of trust calibration. Equivalently, when overtrusting drivers decrease their trust in the AV, they are more likely to reach trust calibration. The increase of trust for undertrusting drivers means that after the communication from the AV, drivers were able to use the self-driving capabilities more confidently, which was reflected by the increases of their related observation variables. Likewise, the framework was able to reduce drivers' trust levels if they presented overtrusting behaviors, when the driving context was not favorable to the AV's autonomous operation. The AV communication demanding drivers' attention to the driving task was effective, tending to adjust (i.e., decrease) drivers' behaviors when they overtrusted the AV.

The proposed real-time trust calibration method was inspired by the relationships among situation awareness, risk perception and trust. Previous works reported on the effectiveness of situation awareness and perceived risk to impact drivers' trust in AVs [25]–[28]. We applied different communications styles and messages in an attempt to vary drivers' situation awareness and risk perception in real time. In consequence, we deliberately induced equivalent real-time changes in trust, supporting the drivers to avoid trust miscalibrations by reducing the difference between their trust estimates and the AV's capability references. The main applicability of the proposed trust management framework is to enable AVs to perceive drivers' trusting behaviors and react to them accordingly. Smart ADSs featuring this capability would likely enhance the collaboration between the driver and the AV because it permits the adaptation of attentional resources according to the operational environment and situation.

Our method can be considered a complement to [7] and [22]. The work in [22] supported our insights for the use of eye-tracking-based techniques for real-time trust estimation. In comparison to [7], we used different methods and behavioral variables for trust estimation and extended their ideas to include the trust calibrator and propose our trust management framework.

The limitations of the framework are mostly related to the uncertainty involved in influencing drivers' trust with different messages, which might not be very effective for some drivers. These drivers might need several interactions to be persuaded by the AV. An example is illustrated in Fig. 7, where the driver was encouraged to trust the AV twice before the increase in $\Delta T = T_3 - T_2$ was registered. The spreads of the box plots represented in Fig. 6 suggest that, in less frequent cases, drivers could present an unexpected behavior, not complying with AV's encouraging or warning messages. The lack of a process for customizing the parameters of our framework contributes to this uncertainty. Relying on average model parameters in the trust estimation block can reduce the accuracy of the estimates because the parameters of each driver can be very different from the averages. Therefore, the trust estimation algorithm (and consequently, the management framework) might work more efficiently if adapted to each individual driver. Another limitation is that the capability of the AV was defined by the circuit track difficulty levels only. Other factors can affect AV capability and could be considered, such as those related to vehicular subsystems or to the weather.

V. Conclusion

This letter proposed a framework for managing drivers' trust in AVs in order to avoid trust miscalibration issues. This framework relies on observing drivers' behaviors to estimate their trust levels, comparing it to capabilities of the AVs, and activating different communication styles to encourage undertrusting drivers and warn overtrusting drivers. Our proposed framework has shown to be effective in inducing positive or negative changes on drivers' trust in the AV and, consequently, mitigating trust miscalibration.

Future works could focus on addressing the limitations of our framework. An individualized system identification scheme, able to capture drivers' behavioral parameters for the dynamic model used in the trust estimator, could be included in the trust estimator algorithm. The individualization of model parameters might increase the effectiveness of the method, enabling more accurate trust estimates and faster trust calibrations. Additionally, as we have assumed that the AV can sense the environment and recognize its own capabilities accordingly, future efforts to develop this capabilities assessment could complete our framework.

The proposed trust management framework is applicable to intelligent driving automation systems, providing them the ability to perceive and react to drivers' trusting behaviors, improving their interaction with the AVs, and maximizing their safety and their performance in tasks other than driving.

ACKNOWLEDGMENT

We greatly appreciate the guidance of Victor Paul (U.S. Army CCDC/GVSC) on the study design, and thank Kevin Mallires for programming the AV simulation scripts.

DISTRIBUTION STATEMENT A. Approved for public release; distribution unlimited.

REFERENCES

[1] T. B. Sheridan and R. T. Hennessy, "Research and modeling of supervisory control behavior. Report of a workshop," National Research Council Committee on Human Factors, Washington, DC., Tech. Rep., 1984. [Online]. Available: https://doi.org/10.17226/19376

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

[3] K. Weir, "The dawn of social robots," *Monitor Psychol.*, vol. 49, no. 1, pp. 50–56, 2018.

[4] B. M. Muir, "Trust between humans and machines, and the design of decision aids," *Int. J. Man-Mach. Stud.*, vol. 27, no. 5-6, pp. 527–539, 1987.

[5] J. D. Lee and N. Moray, "Trust, self-confidence, and operators' adaptation to automation," *Int. J. Human-Comput. Stud.*, vol. 40, no. 1, pp. 153–184, 1994.

[6] D. M. West, "Moving forward: Self-driving vehicles in China, Europe, Japan, Korea, and the United States," Center for Technology Innovation at Brookings, Washington, DC, USA, 2016.

[7] K. Akash, W.-L. Hu, N. Jain, and T. Reid, "A classification model for sensing human trust in machines using EEG and GSR," *ACM Trans. Interactive Intell. Syst.*, vol. 8, no. 4, pp. 1–20, 2018.

[8] J. Metcalfe *et al.*, "Building a framework to manage trust in automation," in *Micro- and Nanotechnology Sensors, Systems, and Applications IX*. Bellingham, WA, USA: SPIE, 2017.

[9] H. Azevedo-Sa, S. K. Jayaraman, C. Esterwood, X. J. Yang, L. Robert, and D. Tilbury, "Real-time estimation of drivers trust' in automated driving systems," *Int. J. Social Robot.*, 2020. doi: 10.1007/s12369-020-00694-1.

[10] T. B. Sheridan, T. Vámos, and S. Aida, "Adapting automation to man, culture and society," *Automatica*, vol. 19, no. 6, pp. 605–612, 1983.

[11] J. Lee and N. Moray, "Trust, control strategies and allocation of function in human-machine systems," *Ergonomics*, vol. 35, no. 10, pp. 1243–1270, 1992.

[12] B. M. Muir, "Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems," *Ergonomics*, vol. 37, no. 11, pp. 1905–1922, 1994.

[13] B. M. Muir and N. Moray, "Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation," *Ergonomics*, vol. 39, no. 3, pp. 429–460, 1996.

[14] B. Barber, *The Logic and Limits of Trust*, vol. 96, New Brunswick, NJ, USA: Rutgers University Press, 1983.

[15] J. K. Rempel, J. G. Holmes, and M. P. Zanna, "Trust in close relationships," *J. Personality Social Psychol.*, vol. 49, no. 1, pp. 95–112, 1985.

[16] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *Int. J. Cognitive Ergonom.*, vol. 4, no. 1, pp. 53–71, 2000.

[17] H. Saeidi *et al.*, "Trust-based mixed-initiative teleoperation of mobile robots," in *Proc. Amer. Control Conf.*, 2016, pp. 6177–6182.

[18] H. Saeidi and Y. Wang, "Incorporating trust and self-confidence analysis in the guidance and control of (semi) autonomous mobile robotic systems," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 239–246, Apr. 2018.

[19] K. Hoff and M. Bashir, "A theoretical model for trust in automated systems," in *CHI '13 Extended Abstracts on Human Factors in Computing Systems*. New York, NY, USA: ACM Press, 2013.

[20] W.-L. Hu, K. Akash, T. Reid, and N. Jain, "Computational modeling of the dynamics of human trust during human-machine interactions," *IEEE Trans. Human-Mach. Syst.*, vol. 49, no. 6, pp. 485–497, Dec. 2019.

[21] C. Castelfranchi and R. Falcone, *Trust Theory: A Socio-Cognitive and Computational Model*. Hoboken, NJ, USA: Wiley, 2010.

[22] Y. Lu and N. Sarter, "Eye tracking: A process-oriented method for inferring trust in automation as a function of priming and system reliability," *IEEE Trans. Human-Mach. Syst.*, vol. 49, no. 6, pp. 560–568, Dec. 2019.

[23] A. R. Marathe *et al.*, "The privileged sensing framework: A principled approach to improved human-autonomy integration," *Theor. Issues Ergonom. Sci.*, vol. 19, no. 3, pp. 283–320, 2018.

[24] L. Petersen, D. Tilbury, X. J. Yang, and L. Robert, "Effects of augmented situational awareness on driver trust in semi-autonomous vehicle operation," in *Proc. Ground Vehicle Syst. Eng. Technol. Symp.*, 2017, pp. 1–7.

[25] L. Petersen, L. Robert, J. Yang, and D. Tilbury, "Situational awareness, driver's trust in automated driving systems and secondary task performance," *SAE Int. J. Connected Auton. Vehicles, Forthcoming*, vol. 2, no. 2, pp. 129–141, 2019.

[26] L. Petersen *et al.*, "The influence of risk on driver's trust in semi-autonomous driving," in *Proc. Ground Vehicle Syst. Eng. Technol. Symp.*, 2018, pp. 1–10.

[27] H. Zhao, H. Azevedo-Sa, C. Esterwood, X. J. Yang, L. Robert, and D. Tilbury, "Error type, risk, performance and trust: Investigating the impacts of false alarms and misses on trust and performance," in *Proc. Ground Vehicle Syst. Eng. Technol. Symp. (GVSETS 2019)*, 2019, pp. 1–9.

[28] H. Azevedo-Sa, S. Jayaraman, C. Esterwood, X. J. Yang, L. Robert, and D. Tilbury, "Comparing the effects of false alarms and misses on humans' trust in (semi) autonomous vehicles," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interact.*, 2020, pp. 113–115.

[29] SAE International, *Taxonomy and Definitions for Terms Related to Driving Automation Systems for on-Road Motor Vehicles*. Warrendale, PA, USA: SAE International, 2018.

[30] H. J. Seltman, *Experimental Design and Analysis*. Pittsburgh, PA, USA: Carnegie Melon University, 2012. [Online]. Available: http://www.stat.cmu.edu/ hseltman/309/Book/Book.pdf

[31] C. A. Miller, "Trust in adaptive automation: The role of etiquette in tuning trust via analogic and affective methods," in *Proc. 1st Int. Conf. Augmented Cognition*, 2005, pp. 22–27.

[32] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "AirSim: High-fidelity visual and physical simulation for autonomous vehicles," in *Proc. Field Service Robot.*, 2017, pp. 621–635. [Online]. Available: https://arxiv.org/abs/1705.05065

[33] S. T. Mueller and B. J. Piper, "The psychology experiment building language (PEBL) and PEBL test battery," *J. Neuroscience Methods*, vol. 222, pp. 250–259, 2014.