Real-Time Estimation of Drivers' Trust in Automated Driving Systems

Hebert Azevedo-Sa · Suresh Kumaar Jayaraman · Connor T. Esterwood · Xi Jessie Yang · Lionel P. Robert Jr · Dawn M. Tilbury

Preprint. Accepted for publication in the *International Journal of Social Robotics*. DOI:10.1007/s12369-020-00694-1

Abstract Trust miscalibration issues, represented by undertrust and overtrust, hinder the interaction between drivers and self-driving vehicles. A modern challenge for automotive engineers is to avoid these trust miscalibration issues through the development of techniques for measuring drivers' trust in the automated driving system during real-time applications execution. One possible approach for measuring trust is through modeling its dynamics and subsequently applying classical state estimation methods. This paper proposes a framework for modeling the dynamics of drivers' trust in automated driving systems and also for estimating these varying trust levels. The estimation method integrates sensed behaviors (from the driver) through a Kalman filter-based approach. The sensed behaviors include eye-tracking signals, the usage time of the system, and drivers' performance on a non-driving-related task (NDRT). We conducted a study ($n = 80$) with a simulated SAE level 3 automated driving system, and analyzed the factors that impacted drivers' trust in the system. Data from the user study were also used for the identification of the trust model parameters. Results show that the proposed approach was successful in computing trust estimates over successive interactions between the driver and the automated driving system. These results encourage the use of strategies for modeling and estimating trust in automated driving systems. Such trust measurement technique paves a path for the design of trust-aware automated driving systems capable of changing their behaviors to control drivers' trust levels to mitigate both undertrust and overtrust.

Keywords Trust · Trust models · Human-robot interaction (HRI) · Automated driving systems · Driving simulation

H. Azevedo-Sa
Robotics Institute, University of Michigan
Ann Arbor, MI 48109, USA
E-mail: azevedo@umich.edu

S. K. Jayaraman
Department of Mechanical Engineering, University of Michigan
Ann Arbor, MI 48109, USA
E-mail: jskumaar@umich.edu

C. T. Esterwood
School of Information, University of Michigan
Ann Arbor, MI 48109, USA
E-mail: cte@umich.edu

X. J. Yang
Robotics Institute, University of Michigan
Ann Arbor, MI 48109, USA
E-mail: xijyang@umich.edu

L. P. Robert Jr
Robotics Institute, University of Michigan
Ann Arbor, MI 48109, USA
E-mail: lprobert@umich.edu

D. M. Tilbury
Robotics Institute, University of Michigan
Ann Arbor, MI 48109, USA
E-mail: tilbury@umich.edu

1 Introduction

Trust is fundamental to effective collaboration between humans and robotic systems [38]. Trust has been studied by the human-robot interaction (HRI) community, especially from researchers who are interested in robotic technologies acceptance and human-robot teams [8, 19, 38, 40, 51]. Researchers have been trying to understand the impacts of robots' behaviors on humans' trust evolution over time [41]. Moreover, they aim to use this understanding to design robots that are aware of humans'

trust to operate in contexts involving collaboration with those humans [7, 8]. Particularly for self-driving vehicles and automated driving systems (ADSs), trust has been used to explore consumer attitudes and enrich the discussion about safety perception [21]. Trust in ADSs, is directly linked to perceptions of their safety and performance which is vital for promoting their acceptance [28, 46, 53].

Trust is a highly abstract concept, and this abstractness makes measuring trust a challenging task [24]. Popular measures of trust are typically self-reported Likert scales, based on subjective ratings. For example, individuals are asked to rate their degree of trust on a scale ranging from 1 to 7 [7, 15, 30]. Although self-reports are a direct way to measure trust, they also have several limitations. First, self-reporting is affected by peoples' individual biases, which makes a precise trust quantification hard to achieve [32]. Second, it is difficult to obtain repeated and updated measures of trust over time without stopping or at least interrupting the task or activity someone is engaged in [10, 52]. Specifically, it is not reasonable to expect ADSs to repeatedly interrupt drivers and ask them to complete a trust surveys. As such, self-reported measures of trust are not an approach that can be relied on to assess drivers' trust in real-time.

An alternative approach to measuring drivers' trust through Likert scale surveys is real-time estimation, done through observing drivers' actions and behaviors. However, there is still much to learn about real-time trust estimation techniques as the current approaches have various limitations. Current approaches fail to provide trust measurements in scales traditionally used for trust in automation [1], or require prohibitive sophisticated sensing and perception methods [1, 25]. These sophisticated methods include the processing of psychophysiological signals (e.g.: galvanic skin response), that are not practical for the vehicular environments, where driver-ADS interactions are likely to take place.

Considering the potential implications for ADS and the far-reaching importance of trust estimation to HRI researchers, our lack of knowledge in this area is a significant gap. For example, given the difficulties involved in measuring real-time trust in the HRI area, such techniques could prove to be valuable across a wide range of robotic interactions with humans. In the case of self-driving vehicles, the ability to indirectly measure trust would open several design possibilities, especially for adaptive ADSs capable of conforming to drivers' trust levels and modifying their own behaviors accordingly. Trust estimations could be used in solutions for issues related to trust miscalibration—i.e., when drivers' trust in the ADS is not aligned with system's actual capabilities or reliability levels [23, 30, 43]. In a simplified approach, trust can be inferred with only the identification and processing of observable variables that may be measured and processed to indicate trust levels. These observation variables essentially represent the behavioral cues present in interactions between drivers and ADSs. However, because of the uncertainty involved in humans' behaviors and actions, a successful trust estimation method must be robust to the uncertainty present in measurements of these observation variables. Predictive models for the variable to be estimated can be used for the development of estimation methods that are robust to uncertainty. Thus, there is a fundamental need for trust dynamic models, describing: (i) how drivers' trust in the ADS changes over time and (ii) the factors that induce changes in drivers' trust in the ADS. This need highlights the importance of developing descriptive models for trust dynamics over the events that occur within driver-ADS interactions. Ultimately, these trust dynamics models are useful for the development of reliable trust estimation techniques.

To address this gap, this paper proposes a framework for the estimation of drivers' trust in ADSs in real-time. The framework is based on observable measures of drivers' behaviors and trust dynamic models. Although different trust estimation approaches have been previously reported in the literature [1, 25], our method is simpler to implement. Those previous approaches represented trust as conditional probabilities. Our trust estimates, instead, are represented in a continuous numerical scale, which is more consistent with Muir's scale [31] and, therefore, also more consistent with the theoretical background on trust in automation. Moreover, our estimation framework relies on a discrete, linear time-invariant (LTI) state-space dynamic model and on a Kalman filter-based estimation algorithm. This formulation makes our trust estimation framework appropriate for treating the unpredictability that characterizes drivers' behaviors and for the design of innovative trust controllers. The trust dynamic model is derived from experimental data obtained in a user experiment with a self-driving vehicle simulator. The estimation algorithm processes observation variables that are suitable for the driver-ADS interaction conditions. This trust estimator is intended to provide a means for the self-driving vehicle's ADS to track drivers' trust levels over time. It enables tracking drivers' trust levels without the need for directly demanding drivers to provide self-reports, which can be disruptive and impractical [24].

The remainder of this paper is organized as follows: Section 2 discusses relevant literature. Sections 3 and 4 establish the theoretical basis for the development of

our model and estimation solution. Section 5 presents details about the user experiment. Section 6 presents the analysis of factors that impact trust and the procedure for trust estimation. Sections 7 and 8 discusses the results and concludes the paper.

2 Related Work

2.1 Trust in Automation and Trust in Robots

Trust in automation has been discussed by researchers since it was first identified as a vital factor in supervisory control systems [39]. Formal definitions of trust in machines came from interpersonal trust theories [3, 33] and were established by Muir in the late eighties [30]. Muir identified the need to avoid miscalibrations of trust in decision aids "so that [the user] neither underestimates nor overestimates its capabilities" [30]. Her work was then extended by Lee and Moray, who used an autoregressive moving average vector form (ARMAV) analysis to derive a transfer function for trust in a simulated semi-automatic pasteurization plant [20]. The inputs for this model were system performance (based on the plant's efficiency) and faults. They later focused on function allocation problems, and found that the difference between trust and self-confidence is crucial for users to define their allocation strategies [22].

The theoretical background on trust in automation has formed the basis for the development of more specific *trust in robots* measurement scales. Schaefer developed a scale that relies on the assessment of forty trust items, related to the human, the robot and the environment where they operate [38]. Yagoda [51] created a measurement scale considering military applications and defining a list of HRI-related dimensions suggested by experts with extensive experience in the field. Charalambous *et al.* gathered qualitative trust-related questions focusing on the industrial human-robot collaboration (HRC) niche, and developed a trust measurement scale for that specific context [6].

In this paper, we consider the widely accepted definition of trust as "*the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability*" [23]. This definition aligns with Muir's standard questionnaire for trust self-reporting, which we used for trust quantification. Trust in automation is distinct from reliance on automation. Trust is an attitude that influences human's reliance behavior, characterized by engaging in automation usage. Trust miscalibrations are likely to induce inappropriate reliance, such as automation misuse or disuse [23].

2.2 Dynamics of Trust and Trust Estimation

Castelfranchi and Falcone [5] define the main aspects of trust dynamics as: how do the experiences of the *trustor agent* (both positive and negative experiences) influence trust changes; and how the instantaneous level of trust influences its subsequent change. These aspects are especially important when a human agent (in this case, the *trustor*) interacts with a machine (i.e., the *trustee*). As in a dynamic system, trust evolution is assumed to depend on the trust condition at a time instance and on the following inputs represented by the trustor's experiences with the trustee [20]. Several works have considered these basic assumptions and presented different approaches for trust dynamics modeling. The argument-based probabilistic trust (APT) model establishes the representation of trust as the probability of a reliable action, given the situation and system features [9]. In the reliance model, reliance is considered a behavior that is influenced by trust [23]. The three-layer hierarchical model describes trust as a result of dispositional, situational and learned factors involved in the human-automation interaction [15].

A relevant approach for modeling the dynamics of trust is that of Hu *et al.* [16], who developed a linear state-space model for the probability of trust responses within two possible choices: trust or distrust in a virtual obstacle detection system. In addition to developing trust-related dynamic models, researchers have tried to use different psychophysiological signals to estimate trust. For instance, extending Hu's work [16], Akash *et al.* [1] proposed schemes for controlling users' trust levels, applying electroencephalography and galvanic skin response measurements for trust estimation. However, psychophysiology-based methods suffer from at least two drawbacks. First and foremost, when using the reported psychophysiological methods, trust is not directly measured. Rather, the results of that method are conditional probabilities of achieving two states (trust or distrust), given prior signal patterns. Although this is a reasonable approach, previous research suggests that trust should be directly measured and represented in a continuous scale [6, 18, 31, 38]. Second, the sensor apparatus applied in psychophysiology-based methods is intrusive and can influence users' performance negatively, bringing practical implementation issues in applications such as self-driving vehicles.

The work presented in this paper differs from previous research in two ways. First, we propose a model that has trust as a continuous state variable, defined in a numerical scale consistent with Muir's subjective scale [31]. Second, we propose a simpler trust sensing method that relies only on eye-tracking as a direct mea-

sure of drivers' behavior. Other variables that are used for sensing are intrinsic to the integration between ADS and the non-driving-related task (NDRT) executed by the driver.

2.3 System Malfunctions and Trust

When not working properly, machines that are used to identify and diagnose hazardous situations—which might trigger human intervention—can present two distinct malfunction types: false alarms and misses [42]. False alarms occur when the system wrongfully diagnoses nonexistent hazards. On the other hand, when the system can not identify the existence of a hazard and no alarm is raised, a miss occurs. These different error types influence system users differently [2,26,27,54], and also have distinct impacts on trust. The influence of false alarms and misses on operators' behaviors was investigated by Dixon *et al.* [12], who has established a relationship with users compliance and reliance behaviors. After being exposed to false alarms, users reduced their compliance behavior, delaying their response to or even ignoring alerts from the system (the "cry wolf" effect). On the contrary, after misses, users allocated more attention to the task environment [11, 47, 48].

It is clear that false alarms and misses represent experiences that influence drivers' trust in ADSs. As systems that are designed to switch vehicle control with the driver in specific situations, ADSs rely on collision sensors that monitor the environment to make the decision to request drivers' intervention. Therefore, while other performance-related factors—such as the ADS's driving styles [4] or failures on different components of the ADS—could affect drivers' trust, we consider that those collision sensors were the most relevant and safety critical elements in SAE level 3 ADSs. In our study, we introduce system malfunctions only in the form of false alarms and misses on the simulated vehicle's collision warning system, while keeping other factors such as the vehicles driving style and other failure types unchanged and generally acceptable: the vehicle followed the standard speed of the road, and no other type of system failure occurred.

3 Problem Statement

Our problem is to estimate drivers' trust in ADS from drivers' behaviors and actions in real-time, while they operate a vehicle equipped with a SAE Level 3 ADS and concurrently perform a visually demanding NDRT. Our method must provide continuous trust estimates that can vary over time, capturing the dynamic nature of drivers' trust in the ADS. The estimation method must avoid the impractical process of repeatedly asking drivers their levels of trust in the ADS, and be as unobtrusive as possible for sensing drivers' behaviors and actions.

4 Method

4.1 Scope

To define the scope of our problem, we make the following assumptions about the ADS and the driving situation:

(i) the ADS explicitly interacts with the driver in events that occur during vehicle operation, and provides automated lane keeping, cruise speed control and collision avoidance capabilities to the vehicle;
(ii) the NDRT device is integrated with the ADS, allowing the ADS to monitor drivers' NDRT performance. The ADS can also track driver's head and eyes orientations;
(iii) drivers can alternate between using and not using the driving automation functions (i.e., the vehicle's self-driving capabilities) at any time during the operation;
(iv) when not using the driving automation functions, drivers have to perform the driving task, and therefore operate the vehicle in regular (non-automated) mode;
(v) using the capabilities provided by the ADS, the vehicle autonomously drives itself when the road is free but it is not able to maneuver around obstacles (i.e., abandoned vehicles) on the road. Instead, the ADS warns the driver whenever an obstacle is detected by the forward collision alarm system, at a fair reaction distance. In these situations, drivers must take over driving control from the ADS and maneuver around the obstacle manually to avoid a collision; and
(vi) the forward collision alarm system is not perfectly reliable, meaning that both false alarms and misses can occur, and the ADS acknowledges when these errors occur. These false alarms and misses lead to interactions that are likely to decrease drivers' trust in the ADS. As mentioned in Section 2.3, no other system malfunctions were implemented in the simulation.

4.2 Problem Solving Approach

Assuming that the variations of trust caused by the interactions between the driver and the ADS can be

quantified, we decide to apply a classical Kalman filter-based continuous state estimation approach for trust. There are three reasons for applying a Kalman filter-based approach: (i) the fact that the continuous output measures of the estimator could be useful for the design of controllers and decision making algorithms in future applications; (ii) the aforementioned well accepted practice of using continuous numerical estimates for trust in automated systems; and (iii) the difficulties related to the stochasticity of drivers' behaviors, which can be mitigated by the Kalman filter with recurring measurements. Therefore, to represent trust as a state variable we need the mathematical derivation of a state-space model that represents the dynamics of trust. We assume that the dynamics of trust is influenced by the trustor agents' instantaneous level of trust and their experiences over time [5].

The implementation of a Kalman filter requires the definition of observation variables that can be measured and processed in real-time. These observation variables must be related to the variable to be estimated. Therefore, to satisfy the ease of implementation requirements stated on Section 3, we select a set of variables that were easy to sense and suitable for being used in a vehicular spatial configuration. The variables are: (i) the amount of time drivers spent using the autonomous capabilities provided by the ADS, i.e., *ADS usage time ratio*; (ii) the relative amount of time drivers spent focusing on a secondary task (the NDRT), measured with an eye-tracker device, i.e., *focus time ratio* [24]; and (iii) drivers' performance on that same NDRT, i.e., *NDRT performance*. The focus time ratio obtained with the eye tracker is chosen because it is conveniently easy to be measured in a vehicle, and has been shown to be successfully representative of trust metrics [24]. The other variables are chosen because they are assumed to be proportional to trust: the more a driver trusts an ADS, the more s/he will use it; the more a driver trusts the ADS, the better s/he will perform on her/his NDRT.

Finally, to identify the parameters of a model for drivers trust in ADS, we need to obtain a training dataset containing both inputs and their corresponding outputs. The outputs must be represented by drivers' true levels of trust in the ADS, which we can obtain by collecting their self-reports in a controlled user experiment. Therefore, only for the purpose of obtaining this training dataset, we establish a procedure for asking drivers their levels of trust in the ADS.

4.3 Definitions

To implement our solution methodology, we must firstly define the terms that will be used in our formulation.

Definition 1 (Trial)

A *trial* is concluded each time the driver operates the vehicle and reaches the end of a predefined route.

Trials are characterized by their time intervals, limited by the instants they start and end. Denoting these by t_0 and t_f, $t_0 < t_f$, the time interval of a trial is given by $[t_0, t_f] \in \mathbb{R}^+$.

Definition 2 (Event)

An *event*, indexed by a $k \in \mathbb{N} \setminus \{0\}$, is characterized each time the ADS warns **or** fails to warn the driver about an obstacle on the road. Events occur at specific time instances t_k corresponding to k, $t_0 < \cdots < t_k < \cdots < t_f$, when the ADS:

(i) correctly identifies an obstacle on the road and alerts the driver to take over control;
(ii) provides a false alarm to the driver; or
(iii) misses an existent obstacle and does not warn the driver about it.

Definition 3 (Event Signals)

The *event signals* are booleans $L(t_k)$, $F(t_k)$ and $M(t_k)$ corresponding to the event k that indicates whether the event was:

(i) a true alarm, for which $L(t_k) = 1$ and $F(t_k) = M(t_k) = 0$;
(ii) a false alarm, for which $F(t_k) = 1$ and $L(t_k) = M(t_k) = 0$; or
(iii) a miss, for which $M(t_k) = 1$ and $L(t_k) = F(t_k) = 0$.

Definition 4 (Instantaneous Trust in ADS)

Drivers' *instantaneous trust in ADS* at the time instance t, $t_0 \leq t \leq t_f$ is a scalar quantity, denoted by $T(t)$.

$T(t)$ is computed from trust variation self-reports and from questionnaires answered by the driver, adapted from the work by Muir and Moray [31]. We re-scale the numerical range of the survey responses to constrain $T(t) \in [T_{min}, T_{max}]$, and arbitrarily choose $T_{min} = 0$ and $T_{max} = 100$. We also assume that $T(t)$ is immutable between two events, i.e., for $t_k \leq t < t_{k+1}$. We consider $T(t)$ to be our basis for the development of the proposed trust estimator.

Definition 5 (Instantaneous Estimate of Trust in ADS)

The *estimate of trust in ADS* at the time instance t, $t_0 \leq t \leq t_f$ is the output of the trust estimator to be proposed, and is represented by $\hat{T}(t)$. Its associated covariance is denoted by $\hat{\Sigma}_T(t)$.

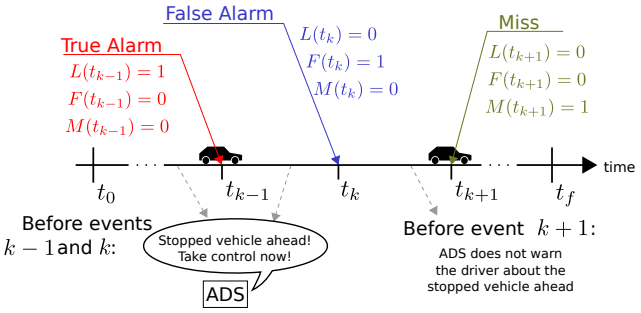


Fig. 1 Timeline example for the stated problem. The event $k-1$ is a true alarm (there is an obstacle car and the ADS warns the driver about it); the event k is a false alarm (there is no car but the ADS also warns the driver); and the event $k+1$ is a miss (there is an obstacle car and the ADS does not warn the driver about it).

Definition 6 (Focus)

Drivers' *focus* on the NDRT, represented by $\varphi(t_k)$, is the percentage of time the driver spends looking at the NDRT screen during the interval $[t_k, t_{k+1})$.

Definition 7 (ADS Usage)

Drivers' *ADS usage*, represented by $\upsilon(t_k)$, is defined by the percentage of time the driver spends using the ADS self-driving capabilities during the interval $[t_k, t_{k+1})$.

Definition 8 (NDRT Performance)

Drivers' *NDRT performance*, represented by $\pi(t_k)$, is the total points obtained by the driver in the NDRT during the interval $[t_k, t_{k+1})$ divided by $\Delta t_k = t_{k+1} - t_k$.

We also call $\varphi(t_k)$, $\upsilon(t_k)$, and $\pi(t_k)$ our *observation variables*.

Fig. 1 shows a timeline scale that represents events within a trial. The NDRT and its score policies are explained in Section 5.

4.4 Trust Dynamics Model

To translate Castelfranchi's and Falcone's main aspects of trust dynamics [5] into mathematical terms, we must represent the experiences of the trustor agent, the subsequent change in trust, and relate those variables. Describing the user experiences with the passing time and the event signals, while also considering their discrete nature, we can expect a general relationship with the form represented by Equation (1),

$$T(t_{k+1}) = f(t_k, T(t_k), L(t_k), F(t_k), M(t_k)), \tag{1}$$

where $f : [t_0, t_f] \times [T_{min}, T_{max}] \times \{0,1\}^3 \to [T_{min}, T_{max}]$.

Additionally, we can expect the relationship between observations and trust to take the form represented by Equation (2),

$$\begin{bmatrix} \varphi(t_k) \\ \upsilon(t_k) \\ \pi(t_k) \end{bmatrix} = h(t_k, T(t_k), L(t_k), F(t_k), M(t_k)), \tag{2}$$

where $h : [t_0, t_f] \times [T_{min}, T_{max}] \times \{0,1\}^3 \to [0,1]^2 \times \mathbb{R}$.

For simplicity, we assume the functions f and h to be linear, time-invariant, with additional random terms representing drivers' individual biases. Moreover, we model trust and the observation variables as Gaussian variables, and consider the observations to be independent of the event signals and within each other, representing the dynamics of trust in the ADS with the LTI system state-space model in Equations (3),

$$\begin{cases} T(t_{k+1}) = \mathbf{A}T(t_k) + \mathbf{B}\begin{bmatrix} L(t_k) \\ F(t_k) \\ M(t_k) \end{bmatrix} + u(t_k); \\ \begin{bmatrix} \varphi(t_k) \\ \upsilon(t_k) \\ \pi(t_k) \end{bmatrix} = \mathbf{C}T(t_k) + w(t_k), \end{cases} \tag{3}$$

where $\mathbf{A} = \begin{bmatrix} a_{11} \end{bmatrix} \in \mathbb{R}^{1\times1}$, $\mathbf{B} = \begin{bmatrix} b_{11}\ b_{12}\ b_{13} \end{bmatrix} \in \mathbb{R}^{1\times3}$, $\mathbf{C} = \begin{bmatrix} c_{11}\ c_{21}\ c_{31} \end{bmatrix}^\top \in \mathbb{R}^{3\times1}$, $u(t_k) \sim \mathcal{N}(0, \sigma_u^2)$ and $w(t_k) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_w)$.

4.5 Trust Estimator Design

The state-space structure permits the application of Kalman filter-based techniques for the estimator design. We then propose the procedure presented in Algorithm 1. Fig. 2 shows a block diagram representation of this framework, highlighting the trust estimator role in the interaction between the driver and the ADS.

5 User Study and Data Collection

We reproduced the situation characterized in Section 4 with the use of an ADS simulator. A total of 80 participants were recruited (aged 18-51, $M = 25.0$, $SD = 5.7$, 52 male, 26 female and 2 who preferred not to specify their genders). Participants were recruited via email and printed poster advertising. All regulatory ethical precautions were taken. The research was reviewed and approved by the University of Michigan's Institutional Review Board (IRB).

Algorithm 1 Trust Estimator

procedure TRUST_ESTIMATION($\hat{T}(t_k), \hat{\Sigma}_T(t_k), L(t_k), F(t_k), M(t_k), \varphi(t_k), \upsilon(t_k), \pi(t_k)$)
  if $k = 0$ **then**
    $\hat{T}(t_0) \leftarrow (\mathbf{C}^\top \mathbf{C})^{-1}\mathbf{C}^\top \begin{bmatrix}\varphi(t_0)\\ \upsilon(t_0)\\ \pi(t_0)\end{bmatrix}$
    $\hat{\Sigma}_T(t_0) \leftarrow 1$ ▷ Initializes trust estimate and co-variance
  else
    $K \leftarrow \hat{\Sigma}_T(t_k)\mathbf{C}^\top(\mathbf{C}\hat{\Sigma}_T(t_k)\mathbf{C}^\top + \Sigma_w)^{-1}$ ▷ Measurement update starting with Kalman gain computation
    $\begin{bmatrix}\hat{\varphi}(t_k)\\ \hat{\upsilon}(t_k)\\ \hat{\pi}(t_k)\end{bmatrix} \leftarrow \mathbf{C}\hat{T}(t_k)$
    $\mathbf{v} \leftarrow \begin{bmatrix}\varphi(t_k)\\ \upsilon(t_k)\\ \pi(t_k)\end{bmatrix} - \begin{bmatrix}\hat{\varphi}(t_k)\\ \hat{\upsilon}(t_k)\\ \hat{\pi}(t_k)\end{bmatrix}$ ▷ Innovation
    $T(t_k) \leftarrow \hat{T}(t_k) + K\mathbf{v}$
    $\Sigma_T(t_k) \leftarrow \hat{\Sigma}_T(t_k) - K\mathbf{C}\hat{\Sigma}_T(t_k)$
    $\hat{T}(t_{k+1}) \leftarrow \mathbf{A}T(t_k) + \mathbf{B}\begin{bmatrix}L(t_k)\\ F(t_k)\\ M(t_k)\end{bmatrix}$ ▷ Time Update
    $\hat{\Sigma}_T(t_{k+1}) \leftarrow \mathbf{A}\Sigma_T(t_k)\mathbf{A}^\top + \sigma_u$
  end if
  return $\hat{T}(t_{k+1}), \hat{\Sigma}_T(t_{k+1})$
end procedure

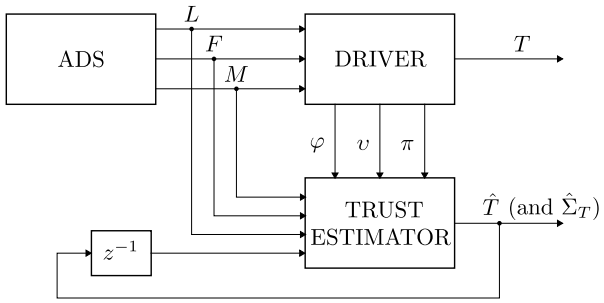


Fig. 2 Block diagram representing the trust estimation framework. The event signals L, F and M indicate the occurrence of a true alarm, a false alarm or a miss. The observations φ, υ and π represent the drivers' behaviors. T is drivers' trust in ADS while $\hat{T}$ and $\hat{\Sigma}_T$ are the estimates of trust in ADS and the covariance of this estimate. A delay of one event is represented by the z^{-1} block.

5.1 Experiment and Data Collection

5.1.1 Study design

We employed a 4 (ADS error types) × 2 (road shapes) mixed user experimental design. Each participant experienced 2 trials, and each trial had 12 events. These 2 trials had the same ADS error type (between-subjects condition) and 2 different road shapes (within-subjects condition). The ADS error types that varied between subjects corresponded to 4 different conditions: control, for which all 12 events were true alarms; false alarms only, for which the 2nd, 3rd, 5th, and 8th events were false alarms; misses only, for which the 2nd, 3rd, 5th, and 8th events were misses; and false alarms and misses combined condition, for which the 2nd and 5th events were false alarms, while the 3rd and 8th events were misses. The ADS error type was assigned according to the participants' sequential identification number. The road shapes were represented by straight and curvy roads, and were assigned in alternating order to minimize learning and ordering effects.

5.1.2 Tasks

We used a driving simulation designed and implemented with the *Autonomous Navigation Virtual Environment Laboratory* (ANVEL) simulator [13]. The NDRT was an adapted version of the Surrogate Reference Task [17], implemented with the *Psychology Experiment Building Language* (PEBL) [29]. Fig. 3(a) shows the experimental setup with the tasks performed by the driver.

In the driving task, participants operated a simulated vehicle equipped with an ADS that provided it automatic lane keeping, cruise control, and collision avoidance features. Participants were able to activate the ADS (starting autonomous driving mode) by pressing a button on the steering wheel, and to take back control by braking or by steering. Fig. 3(b) shows the driving task interface with the driver.

With the ADS activated (i.e., with the vehicle in self-driving mode), participants were expected to execute the visual search NDRT. They were not allowed to engage in both driving and executing the NDRT simultaneously, and the experimenters would stop the test if they did so. Participants were informed that the vehicle could request their intervention if they identified obstacles on the road, as it is expected for Level 3 ADSs [35]. They needed to find a "Q" character among several other "O" characters, and obtained 1 point for each correctly chosen "Q". Fig. 3(c) shows the NDRT interface with the driver.

Participants could not focus only on the NDRT, because the ADS demanded them to occasionally take control of the driving task. They were asked to be ready to take control upon intervention requests from the ADS, as some obstacles occasionally appeared on the road. At that point, the ADS identified the obstacles and asked the driver to take control, as the vehicle was not able to autonomously change lanes and maneuver around them. If drivers did not take control, the emergency brake was triggered when the vehicle got too close to an obstacle, and then drivers lost points on their on-

going NDRT score. In that situation, they still needed to take control of the driving task, maneuver around the obstacle and re-engage the autonomous driving mode. They lost 5 points each time the emergency brake got triggered.

With the events characterized by true alarms or misses, drivers had to take control and pass the obstacle. Subsequently, they were asked about their "trust change". When asked, they had to stop the vehicle to answer the question on a separate touchscreen. They reported their trust change in the events characterized by true alarms, false alarms, and misses. They had 5 choices, varying from "Decreased Significantly" to "Increased Significantly", as shown in Fig. 3(d). These choices were then used as indicators of the differences $\Delta T_k^Q \in \{-2, -1, 0, 1, 2\}$ (we use the superscript Q to indicate that the differences were quantized).

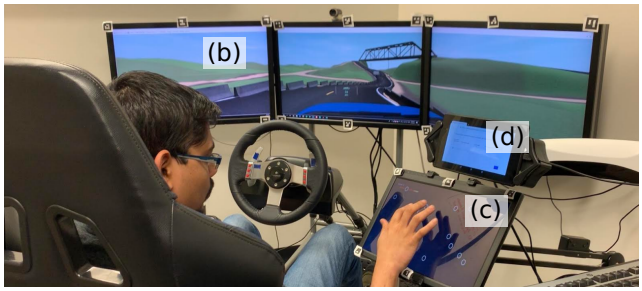
5.1.3 Procedure

Upon arrival, participants were asked to complete a consent form as well as a pre-experiment survey related to their personal information, experience with ADS, mood and propensity to trust the ADS. After the survey, the tasks were explained and the experimenter gave details about the experiment and the simulated vehicle control. Participants then completed a training session before the actual experiment began and, in sequence, completed their two trials. After each trial, participants were asked to complete post-trial surveys related to their trust in the ADS. These surveys were administered electronically. Each trial took approximately 10 to 15 minutes, and the whole experiment lasted approximately 60 minutes.

A basic fixed level of cash compensation of $15.00 was granted for the participants. However, they also had the possibility of receiving a performance bonus. The bonus was calculated according to their best final NDRT score, considering both trials experienced by the participant. Those who made up to 199 points in the NDRT did not receive a bonus. However, bonuses of $5.00 were granted for those who made between 200 and 229 points; $15.00 for those who made between 230 and 249 points; and $35.00 for those who made 250 points or more. From the total of 80 participants, 28 got $5.00 bonuses, 6 participants got $15.00 bonuses, and no participant got the $35.00 bonus.

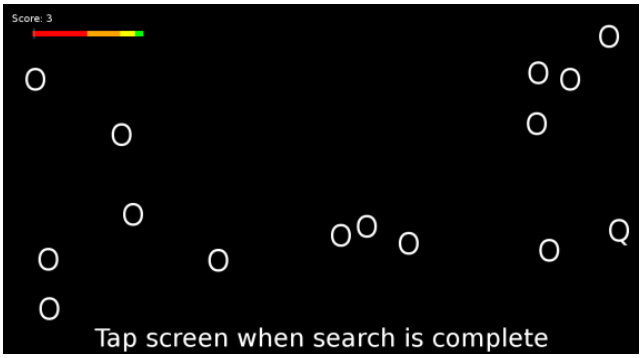
5.1.4 Apparatus

As illustrated in Fig. 3(a), the simulator setup was composed of three LCD monitors integrated with a *Logitech*

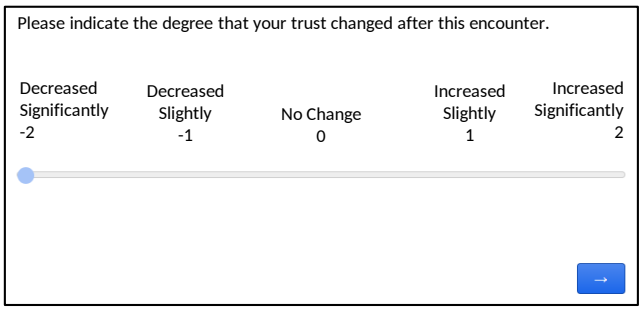


(a)

(b)



(c)



(d)

Fig. 3 Experimental design (a), composed of the driving task (b), the NDRT (c) and the trust change self-report question (d). The trust change self-report question popped up after every event within the trials (there were 12 events per trial), including true alarms, false alarms, and misses.

G-27 driving kit. Two other smaller touchscreen monitors positioned to the right hand of the participants were used for the NDRT and for the trust change self-report questions. The console was placed to face the central monitoring screen so as to create a driving experience as close as possible to that of a real car. In addition, we used *Pupil Lab's Pupil Core* eye tracker mobile headset, equipped with a fixed "world camera" to measure participants' gaze positional data.

5.1.5 Measured Variables

Measured variables included participants' subjective responses, behavioral responses and performance. Observation variables $\varphi(t_k)$, $\upsilon(t_k)$ and $\pi(t_k)$ were also measured and averaged for the intervals $[t_k, t_{k+1}]$. Subjective data was gathered through surveys before and after each trial, including trust perception, risk perception, and workload perception. We used questionnaires adapted from [31] and [34] to measure post-trial trust and risk perception, respectively. Eye-tracking data included eyes' positions and orientations, as well as videos of the participants' fields of view.

$T(t_k)$ was computed from the post-trial trust perception self-reports $T(t_f)$ and the within trial trust change self-reports ΔT_k^Q, as in Equation (4),

$$\begin{cases} T(t_{12}) = T(t_f)\,; \\ T(t_k) = T(t_f) - \alpha \sum_{i=k+1}^{12} \Delta T_i^Q\,, \end{cases} \tag{4}$$

where $k \in \{0, 1, 2, \ldots, 11\}$, and $\alpha = 3$. Therefore, the trust measures $T(t_k)$ were back-computed for the events within a trial. The α value was chosen to characterize noticeable variations in $T(t_k)$, but also avoiding $T(t_k)$ values falling outside the interval $[T_{min}, T_{max}]$. Positive values for α between 1 and 3 were tested and provided results similar to those reported in Section 6.

5.2 Model Parameters

Considering the formulation presented in Section 4 and the data obtained in the user study, we turn to the identification of parameters for the trust model and the design of the trust estimator. We found the best fit parameters for the short-term (i.e., with respect to events) trust dynamics represented by the state-space model in Equation (3). From the 80 participants, we selected 4 from the dataset—each one chosen randomly within each of the 4 possible ADS error type conditions—and used the data from the remaining 76 to compute the parameters, which are presented in Table 1. We used the data from the 4 selected participants for validation.

Table 1 Trust in ADS state-space model parameters

Parameter	Value Estimate	S.E.M†
a_{11}	0.9809	4.0×10^{-3}
b_{11}	3.36	0.29
b_{12}	-0.61	0.32
b_{13}	-1.30	0.31
c_{11}	6.87×10^{-3}	3.3×10^{-4}
c_{21}	9.10×10^{-3}	1.0×10^{-4}
c_{31}	4.38×10^{-3}	1.0×10^{-4}
σ_u^2	1.24	–
$\mathbf{\Sigma}_w$	$\mathrm{diag}(1.0, 1.6, 1.8) \times 10^{-3}$	–

†S.E.M = Standard error of the mean.

The parameters of the state-space model from Equation (3) were identified with maximum likelihood estimation through linear mixed-effects models. Our models included a random offset per participant to capture their individual biases and mitigate the effects of these biases in the results, and to represent normally distributed random noises.

6 Results

6.1 Participants' Data Analysis

For each of the observation variables, we obtained 1920 measurements (80 participants × 2 trials per participant × 12 events per trial). The parameters describing these distributions are presented in Table 2. The histograms for these distributions are shown in Fig. 4; the probability density functions corresponding to normal distributions $\mathcal{N}(\mu_\varphi, \sigma_\varphi^2)$, $\mathcal{N}(\mu_\upsilon, \sigma_\upsilon^2)$ and $\mathcal{N}(\mu_\pi, \sigma_\pi^2)$ are also shown.

Table 2 Parameters for the Focus φ, ADS usage υ and NDRT performance π measurements distributions

Parameter	Distributions		
	φ	υ	π
Minimum	0.02	0.17	0.00
25^{th} percentile	0.32	0.69	0.28
50^{th} percentile	0.47	0.74	0.33
75^{th} percentile	0.65	0.79	0.38
Maximum	0.97	0.92	0.56
Mean μ	0.49	0.73	0.32
Standard Deviation σ	0.20	0.08	0.08

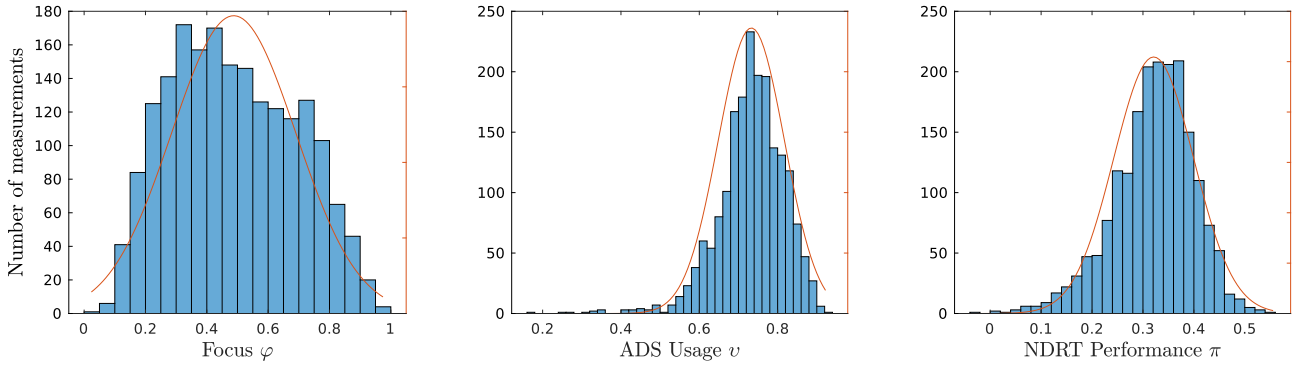


Fig. 4 Histograms for the Focus φ, ADS usage υ and NDRT performance π measurements distributions and overlapping probability density functions with corresponding means and standard deviations. Each distribution had 1920 measurements (= 80 participants × 2 trials per participant × 12 measurements per trial).

6.2 Trust Estimation Results

After obtaining the model parameters, we applied Algorithm 1 to estimate the trust levels of the participants that were excluded from the dataset. Fig. 5(a1:a4) and Fig. 6(a1:a4) present the trust estimation results for these participants (identified as A, B, C and D). Participant A experienced the combined ADS error type condition; participant B experienced the false alarms only condition; participant C experienced the control condition; and participant D experienced the misses only condition. The plots bring together their two trials and the different estimate results for each trial. For participants A and B, trial 1 was conducted on a curvy road and trial 2 on a straight road. For participants C and D, trial 1 was conducted on a straight road and trial 2 on a curvy road.

The accuracy of our estimates improved over time, as the participants interacted with the ADS. Fig. 5(a1) shows that, for participant A, trial 1, the initial trust estimate $\hat{T}(t_0)$ and the initial observed trust $T(t_0)$ were close to each other (in comparison to Fig. 5(a2)). This means that the estimate computed from the observations taken at the beginning of the trial, i.e., $\varphi(t_0)$, $\upsilon(t_0)$, and $\pi(t_0)$, approximately matched the participants self-reported trust level. Considering the Kalman filter's behavior, the curves remained relatively close together over the events, as expected. Therefore the estimate followed the participants' trust over the trial events. This accuracy, however, was not achieved at the beginning of the second trial, as can be observed in Fig. 5(a2). This figure shows that, in trial 2, $\hat{T}(t_0)$ and $T(t_0)$ had a greater difference, but this difference decreased over the events as the curves converged. A similar effect can be observed for participants B, trial 2 as in Fig. 5(a3:a4) and for participant C, as in Fig. 6(a1:a2).

Participants' responses to similar inputs were not always coherent, and varied over time or under certain conditions. Predominantly, participants' self-reported trust increased after true alarms (indicated by the prevailing positive steps at the events that are characterized by orange circles). In addition, after false alarms and misses, they usually reported trust decreases (indicated by the prevailing negative steps at the events characterized by yellow diamonds and purple triangles). However, it is noticeable that, for participant A, trial 2, the self-reported trust was more "stable", as indicated by fewer steps on the red dashed curve. Two different factors could have contributed to the less frequent variations on $T(t_k)$: as the participant was on a straight road, the perceived risk might not have been high enough to induce drops after false alarms; or, as it was the participant's second trial, the learning effects might have softened the self-reported trust changes (especially after false alarms). In any case, the difference between the curve patterns in Fig. 5(a1) and Fig. 5(a2) suggests a non-constancy on participant A's characteristic behaviors. A similar behavior was observed for participant C, trial 1 after the 8^{th} alarm and for trial 2.

The observation variables we selected were effective in representing drivers trusting behaviors. Fig. 5(b1:d4) show the observation variables corresponding to the trust curves in Fig. 5(a1:a4), while Fig. 6(b1:d4) correspond to 6(a1:a4). All observation variables have a positive correlation with trust, and therefore it can be observed that some noticeable peaks and drops in the observation variables correspond to positive and negative variations in the estimate of trust in ADS. This is especially true for counterintuitive behaviors of the participants. For instance, as it can be seen in Fig. 5(a3:d3), after the 8^{th} event—which was a false alarm—participant B reported a drop in his/her trust level,

indicating that $T(t_8) < T(t_7)$. However, his/her behaviors did not reflect that drop: we can notice that $\varphi(t_8) > \varphi(t_7)$, $\upsilon(t_8) > \upsilon(t_7)$ and $\pi(t_8) > \pi(t_7)$. As a result, the trust estimate had an increase, and eventually we had $\hat{T}(t_8) > \hat{T}(t_7)$. Similar counter-intuitive situations can be identified for participants A, C and D.

The accuracy of the estimates depends on the covariance parameters, which can be tailored for the driver. The trust estimate bounds represented by blue bands in Fig. 5(a1:a4) and Fig. 6(a1:a4) are approximations obtained with the overlay of several simulations (100 in total). This variability is due to the uncertainty represented by the random noise parameters $u(t_k)$ and $w(t_k)$, and the width of the bound bands is related to the computed covariances σ_u^2 and $\mathbf{\Sigma}_w$. Both lower values for σ_u^2 and higher values for $\mathbf{\Sigma}_w$ entries would imply a narrower band, meaning that the estimator would have less variability (and therefore could be slower on tracking trust self-reports). Meanwhile, higher σ_u^2 and lower values of $\mathbf{\Sigma}_w$ entries would imply, respectively, a less accurate process model and on observations considered more reliable. This would characterize wider bands, and thus the variations on the estimate curves would be more pronounced.

Trust estimates may be more accurate with the individualization of the model parameters. Although we used the average parameters presented in Table 1 for the results, a comparison of Fig. 5(a2), Fig. 6(a1) and Fig. 6(a3:a4) with Fig. 5(a4), suggests that the balance between σ_u^2 and $\mathbf{\Sigma}_w$ should be adapted to each individual driver. It can be seen that these parameters permitted a quick convergence of $T(t_k)$ and $\hat{T}(t_k)$ for participants A, C and D, but that 12 events were not enough for the estimator to track the trust self-reports from participant B. We also computed the root-mean-square (RMS) error of the estimate curves resulting from the 100 simulations for participants A, B, C and D. The RMS error distributions had the characteristics presented in Table 3.

Considering the 100-points trust range, for participant A the error stands below 10%, while for participants B, C and D it stands below 20%. This difference suggests that the parameters of the model are more suitable for participant A than for participant B, C and D.

Table 3 RMS error of the estimate curves from Fig. 5 and Fig. 6

Participant	Trial	Mean	Standard Deviation
A	1	4.9	2.4
A	2	10.0	2.1
B	1	14.5	2.8
B	2	19.1	1.2
C	1	14.2	0.4
C	2	2.7	0.6
D	1	20.7	2.2
D	2	13.8	3.4

7 Discussion

7.1 Contributions and Implications

The goal of this paper was to propose a framework for real-time estimation of drivers' trust in ADS based on drivers' behaviors and dynamic trust models. As shown by the results, our framework successfully provides estimates of drivers' trust in ADS that increase in accuracy over time. This framework is based on a novel methodology that has considerable advantages over previously reported approaches, mainly related to our trust dynamics model and the simpler methods needed for its implementation.

First, the sensing machinery required for implementing our methodology is as simple and as unobtrusive as possible. Considering practical aspects related to the framework implementation, we have chosen observation variables that are suitable for the estimation of drivers' trust in ADS. An eventual implementation of the proposed estimator on an actual self-driving vehicle would depend only on the utilization of an eye-tracking system and on the integration between the ADS and the tasks performed by the driver. Our unique observation variable that comes from a direct instrumentation of drivers' behavioral patterns is the eye-tracking-based focus on the NDRT. The other observation variables (NDRT performance and ADS usage) are indirectly measured by the ADS. Eye-tracking-based metrics are appropriate for trust measuring as they do not require sensory devices that would be impractical and/or intrusive for drivers. Although we have used an eye tracker device that has to be directly worn by the participant, there exist different eye-tracking systems that do not need to get in direct contact with the driver to sense their gaze orientations, and could be used in a real world implementation of this framework.

Second, the results of our framework show that it can successfully estimate drivers' trust in ADS levels, but the estimates accuracy were different depending on the driver. The application of the model represented by Equation (3) in the trust estimator algorithm required average (population-wise) state-space model parameters. These parameters were computed with a minimization problem, and they are indications of reason-

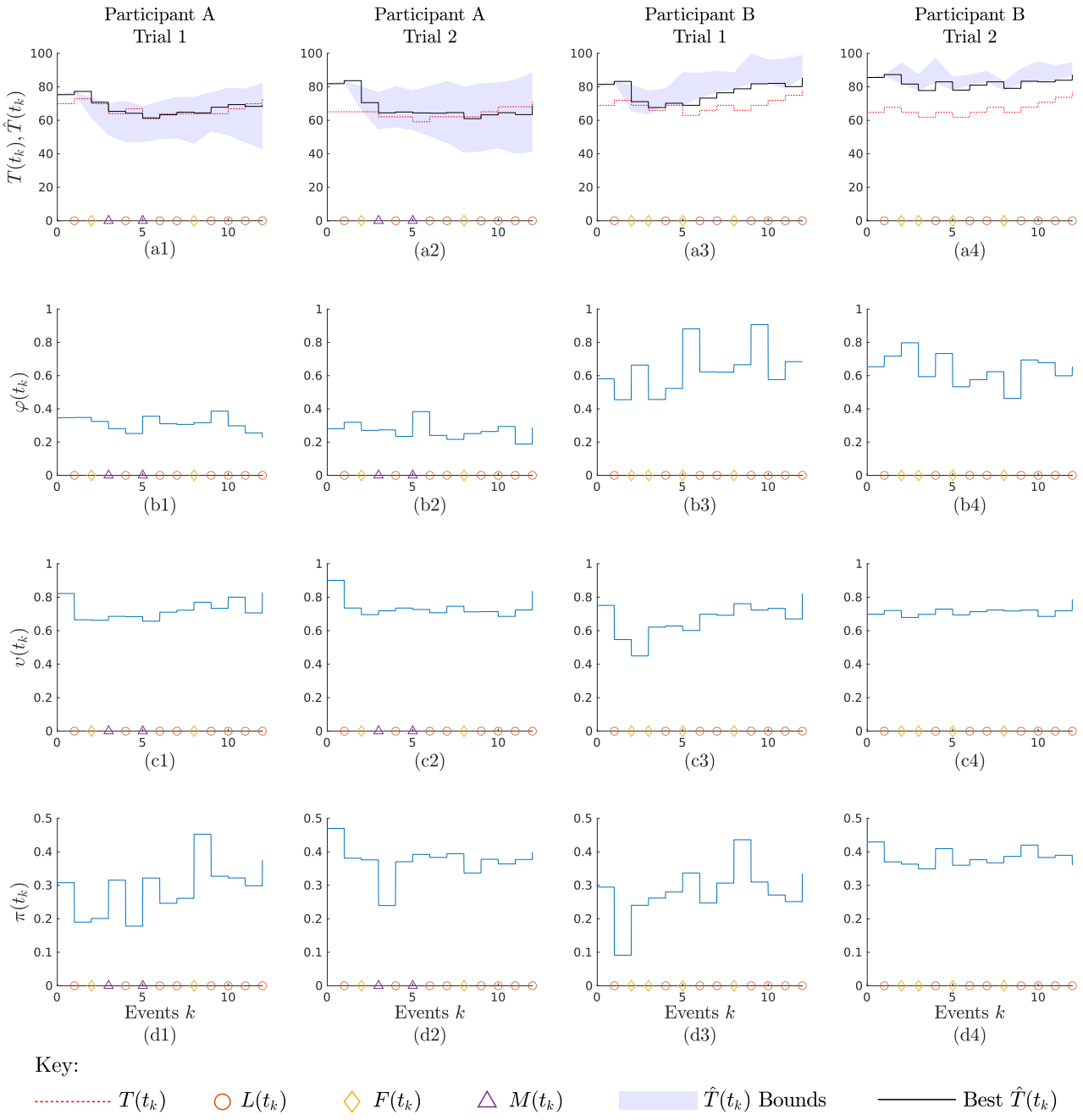


Fig. 5 Trust estimation results for participants A and B. Participant A experienced both false alarms and misses (combined ADS error type condition) while participant B experienced false alarms only (false alarms only condition). For both participants, the first trial was conducted on a curvy road, while the second trial was conducted on a straight road. Curves in (a1:a4) show the estimation results, indicating that the estimator can track the trust self-reports, i.e., $\hat{T}(t_k)$ approaches $T(t_k)$ over the events. This is made possible with the processing of the observations variables focus time ratio (φ), ADS usage time ratio (υ), and NDRT performance (π) presented in (b1:d4).

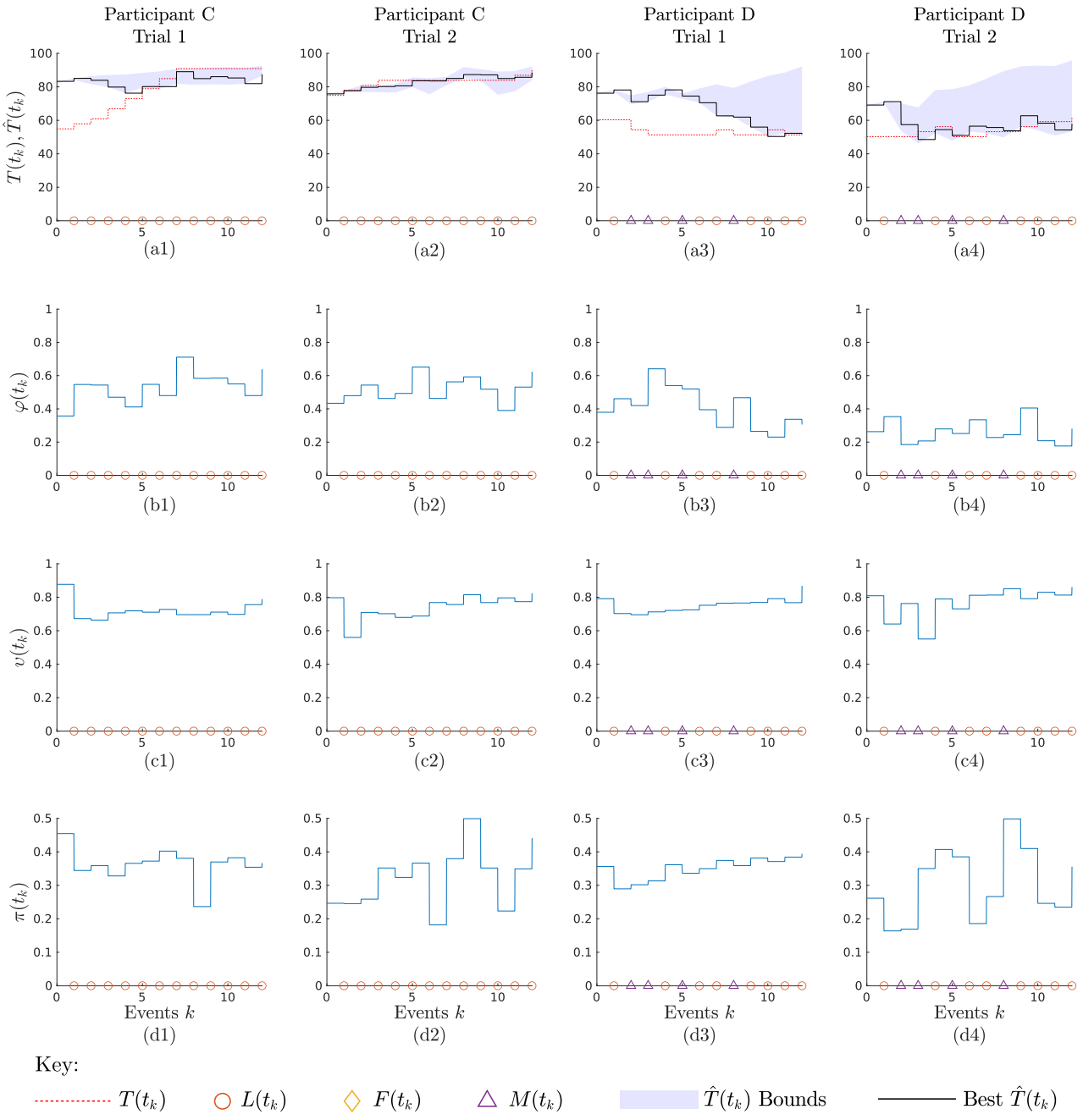


Fig. 6 Trust estimation results for participants C and D. Participant C experienced only true alarms (control ADS error type condition) while participant D experienced misses only (misses only condition). For both participants, the first trial was conducted on a straight road, while the second trial was conducted on a curvy road.

able statistics for average values conditioned to our pool of participants. However, these parameters could vary drastically from driver to driver. In a more sophisticated implementation of our modeling and estimation methodology, the values from Table 1 should serve as preliminary parameters only. A possible way to improve our proposed methodology would be to the integrate it with learning algorithms to adapt the model parameters to individual drivers. Moreover, as drivers become accustomed to the ADS's operation, these parameters might also vary over time (making the time-invariant description from Equation (3) not useful). Therefore, an eventual ADS featuring our framework should also be sufficiently flexible to track the changes in individ-

ual drivers' model parameters over time, as proposed in [49].

Third, the paper's framework opens paths for more research on the development of more complex models and estimation techniques for trust. These techniques may encompass both the driver-ADS context and other contexts characterized by the interaction between humans and robots. In the case of driver-ADS contexts, the events that trigger the propagation of the trust state do not need to be restricted to the forward collision alarm interactions characterized by true alarms, false alarms and misses. A wider range of experiences could be considered in the process model represented by Equation (3), such as events related to the ADS driving performance or to external risk perceived by the ADS. Drivers could be engaged in alternative NDRTs, as long as they are integrated with the ADS and a continuous performance metric is defined as observation variable. In the case of interactions between humans and robots in different scenarios, the concepts that were defined in Section 4 are easily expandable to other contexts. The main requirement would be the characterization of what are the events that represent important (positive and negative) experiences within interactions between the human and robot. These positive and negative experiences would generally characterize the robot's performance, which is an essential factor describing the basis of trust, as identified by Lee and See [23]. Robots that execute specific tasks in goal-oriented contexts could have their performances measured in sequential time instances that would trigger the the transition of the trust state. For instance, these performance measures could be a success/failure classification, such as pick and place task with a robotic arm [40, 44, 50]; or a continuous performance evaluation, such as when a follower robot loses track of its leader due to the accumulation of sensor error [36, 37].

Finally, the paper's framework provides trust estimates that are useful for the design of trust controllers to be embedded in new ADSs. In our framework, trust is modeled as a continuous state variable, which is consistent with widely used trust scales and facilitates the processing and analysis of trust variations over time. This trust representation permits considering the incremental characteristics of the trust development phenomena, which is consistent with the literature on trust in automation and opens a path for the development of future trust control frameworks in ADSs. Since it is developed in the state-space form, our method for modeling drivers' trust in ADS enables the use of classical application-proven techniques such as the Kalman filter-based method we have used in Algorithm 1.

In addition, a practical implication of the proposed estimation framework is that it could be used in innovative adaptive systems capable of estimating drivers' trust levels and reacting in accordance with the estimates, in order to control drivers' trust in ADS. These functionalities would need to involve strategies to monitor not only drivers' behaviors but also the reliability of the system (for example, the acknowledgment of false alarms and misses mentioned in Section 4.1, assumption (vi)). These errors could be identified after a sequence of confirmations or contradictions of the sensors' states, while the vehicle gets closer to the event position, entering the ranges of higher accuracy of those sensors. Moreover, the system could request the driver to provide it feedback about issued alarms to identify its own errors, asking confirmation about identified obstacles or enabling quick report of missed obstacles, a functionality that is currently present in GPS navigation mobile applications [45]. Although these questions could represent an inconvenient distraction, this strategy is not as disruptive as demanding drivers to provide trust self-reports, especially during autonomous operation. The integration between the ADS and the NDRTs would also be needed for the assessment of observation variables and, eventually, actions to increase or decrease trust in ADS could be taken to avoid trust-related issues (such as under- and over-trust). These trust control schemes would be useful for improving driver-ADS interactions, having the goal of optimizing the safety and the performance of the team formed by the driver and the vehicle.

7.2 Limitations

7.2.1 Trust modeling and Estimation Methodology

A limitation of our study relates to the assumptions associated with how we derive the state-space model for trust in the ADS. The relationships represented by Equations (1) and (2) restrict the experiences of the trustor agent (the driver) to the events represented by true alarms, false alarms and misses of the forward collision alarm. In fact, other experiences such as the ADS's continuous driving performances can characterize events that could be represented by signals of different types other than booleans. The simplification of the relationships represented by (1) and (2) to the LTI system represented by (3) is useful and convenient for the system identification process and for the trust estimator design. However, the resulting model fails to capture some phenomena that are likely to occur during the interactions between drivers and ADSs. These

phenomena might include the variation of model parameters over time (i.e., after a reasonable period of drivers' interaction with the ADS) or the possibly nonlinear relationship between trust and the observation variables. An example is the relationship between trust and NDRT performance: it is unlikely that in a more rigorous modeling approach we could consider these variables to be directly proportional. Usually an excess of trust (overtrust) in a system can lead to human errors, which might eventually result in performance drops.

7.2.2 User Study

There are several other limitations that relate to our experimental study. First, most participants were young students, very experienced with video games and other similar technologies. Our results could have been biased by these demographic characteristics.

Second, we employed a simulator in our experimental study. The use of a simulated driving environment is a means of testing potentially dangerous technologies. In general, people tend to act similarly in real and simulated environments [14]. However, due to the risks involved in driving, we acknowledge that participants might not have felt as vulnerable as they would if this study had been conducted in a real car.

Finally, we employed a specific NDRT to increase the participants' cognitive load. The recursive visual search task gives drivers the opportunity to switch their attention between the driving and the NDRT very frequently. Other types of NDRTs could demand drivers' attention for longer periods of time, and this could induce a different effect on trust, risk perception or performance. The NDRT performance metric in this study is very specific and may or may not be generalizable to other task types.

7.3 Future Work

Future research should focus on the use of this modeling technique to design a trust management system composed of the estimator and a trust controller. The trust management system could compare the trust level estimates with the assessed capability and reliability of the vehicle in different situations, which would depend on the risk involved in the operation. From the comparison, the trust calibration status could be evaluated, and a possible mismatch between trust and capability (or reliability) levels would indicate the need for system reaction. This reaction would consist of actions to manipulate trust levels, seeking to increase trust in case of distrust (or undertrust) and to decrease it in case of overtrust.

Additional improvements to our framework may be achieved by addressing the limitations of the reported user study. A vehicle with autonomous capabilities can be utilized to make the participants' experience as similar as possible to a realistic situation. Additionally, our methodology could be tested in other different scenarios where the complexity of the NDRT and of the environment are increased.

8 Conclusion

In this paper we presented a framework for the estimation of drivers' trust in ADSs. Our framework is applicable for SAE level 3 ADSs, where drivers conditionally share driving control with the system, and that system is integrated with a visually demanding NDRT. In comparison to previous trust estimation approaches, it has practical advantages in terms of implementation ease and of the format of its trust estimates outputs.

We investigated the effectiveness of the proposed framework with a user study that is reported in Section 5. In this user study, participants operated a simulated vehicle featuring an ADS that provided self-driving capabilities for the vehicle. Participants conducted two concurrent (driving and non-driving) tasks, while reporting their levels of trust in the ADS. Our goal was to establish a computational model for drivers' trust in ADS that permitted trust prediction during the interactions between drivers and ADSs, considering the behaviors of both the system and the driver. We found the parameters of a discrete-time, LTI state-space model for trust in ADS. These parameters represented the average characteristics of our drivers, considering the resultant experiment dataset. With the parameters calculation it was possible to establish a real-time trust estimator, which was able to track the trust levels over the interactions between the drivers and the ADS.

In summary, our results reveal that our framework was effective for estimating drivers' trust in ADS through the integration of the NDRT and behavioral sensors to ADSs. We also show, however, that a more advanced strategy for trust estimation must take into consideration the individual characteristics of the drivers, making systems flexible enough to adjust their model parameters during continuous use. Our technique opens ways for the design of smart ADSs able to monitor and dynamically adapt their behaviors to the driver, in order control drivers' trust levels and improve driver-ADS teaming. More accurate trust models can improve the performance of the proposed trust estimation framework and, therefore, are still required. However, the

utilization of this trust estimation framework can be a first step to designing systems that can, eventually, increase safety and optimize joint performances during the interactions between drivers and ADSs embedded in self-driving vehicles.

Acknowledgements We greatly appreciate the guidance of Mr. Victor Paul in helping with the study design. The authors would also like to thank Quantum Signal, LLC, for providing ANVEL software and their development support.

Distribution Statement

DISTRIBUTION A. Approved for public release; distribution unlimited. OPSEC #:3816

Compliance with Ethical Standards

Funding: This research is partially supported by the National Science Foundation, the Brazilian Army's Department of Science and Technology and the Automotive Research Center (ARC) at the University of Michigan, through the U.S. Army CCDC/GVSC (government contract DoD-DoA W56HZV14-2-0001).
Conflict of interest: The authors declare that they have no conflict of interest.

References

1. Akash, K., Hu, W.L., Jain, N., Reid, T.: A Classification Model for Sensing Human Trust in Machines Using EEG and GSR. ACM Transactions on Interactive Intelligent Systems **8**(4), 1–20 (2018). DOI 10.1145/3132743. URL `http://dl.acm.org/citation.cfm?doid=3292532.3132743`
2. Azevedo-Sa, H., Jayaraman, S., Esterwood, C., Yang, X.J., Robert, L., Tilbury, D.: Comparing the effects of false alarms and misses on humans' trust in (semi) autonomous vehicles. In: 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI). ACM (2020). DOI 10.1145/3371382.3378371
3. Barber, B.: The logic and limits of trust, vol. 96. Rutgers University Press New Brunswick, NJ (1983)
4. Basu, C., Yang, Q., Hungerman, D., Sinahal, M., Draqan, A.D.: Do you want your autonomous car to drive like you? In: 2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI, pp. 417–425. IEEE (2017)
5. Castelfranchi, C., Falcone, R.: Trust theory : a socio-cognitive and computational model. John Wiley & Sons, West Sussex, United Kingdom (2010). URL `http://cds.cern.ch/record/1319739`
6. Charalambous, G., Fletcher, S., Webb, P.: The development of a scale to evaluate trust in industrial human-robot collaboration. International Journal of Social Robotics **8**(2), 193–209 (2016)
7. Chen, M., Nikolaidis, S., Soh, H., Hsu, D., Srinivasa, S.: Planning with trust for human-robot collaboration. In: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, pp. 307–315 (2018)
8. Chen, M., Nikolaidis, S., Soh, H., Hsu, D., Srinivasa, S.: Trust-aware decision making for human-robot collaboration: Model learning and planning. ACM Transactions on Human-Robot Interaction (THRI) **9**(2), 1–23 (2020)
9. Cohen, M.S., Parasuraman, R., Freeman, J.T.: Trust in decision aids: A model and its training implications. In: in Proceedings of Command and Control Research and Technology Symposyum, pp. 1–37. Cognitive Technologies, Arlington, VA (1998)
10. Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., Yanco, H.: Impact of robot failures and feedback on real-time trust. In: 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 251–258. IEEE (2013)
11. Dixon, S.R., Wickens, C.D.: Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. Human factors **48**(3), 474–486 (2006)
12. Dixon, S.R., Wickens, C.D., McCarley, J.S.: On the independence of compliance and reliance: Are automation false alarms worse than misses? Human Factors **49**(4), 564–572 (2007). DOI 10.1518/001872007X215656. URL `http://journals.sagepub.com/doi/10.1518/001872007X215656`
13. Durst, P.J., Goodin, C., Cummins, C., Gates, B., Mckinley, B., George, T., Rohde, M.M., Toschlog, M.A., Crawford, J.: A real-time, interactive simulation environment for unmanned ground vehicles: The autonomous navigation virtual environment laboratory (anvel). In: 2012 Fifth International Conference on Information and Computing Science, pp. 7–10. IEEE, Shanghai, China (2012)
14. Heydarian, A., Carneiro, J.P., Gerber, D., Becerik-Gerber, B., Hayes, T., Wood, W.: Immersive virtual environments versus physical built environments: A benchmarking study for building design and user-built environment explorations. Automation in Construction **54**, 116–126 (2015)
15. Hoff, K., Bashir, M.: A theoretical model for trust in automated systems. In: CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13, p. 115. ACM Press, New York, New York, USA (2013). DOI 10.1145/2468356.2468378. URL `http://dl.acm.org/citation.cfm?doid=2468356.2468378`
16. Hu, W.L., Akash, K., Reid, T., Jain, N.: Computational Modeling of the Dynamics of Human Trust During Human-Machine Interactions. IEEE Transactions on Human-Machine Systems **1**(1), 1–13 (2018). DOI 10.1109/THMS.2018.2874188. URL `https://ieeexplore.ieee.org/document/8502860/`
17. Jamson, A.H., Merat, N.: Surrogate in-vehicle information systems and driver behaviour: Effects of visual and cognitive load in simulated rural driving. Transportation Research Part F: Traffic Psychology and Behaviour **8**(2), 79–96 (2005)
18. Jian, J.Y., Bisantz, A.M., Drury, C.G.: Foundations for an empirically determined scale of trust in automated systems. International journal of cognitive ergonomics **4**(1), 53–71 (2000)
19. Kessler, T.T., Larios, C., Walker, T., Yerdon, V., Hancock, P.: A comparison of trust measures in human–robot interaction scenarios. In: Advances in Human Factors in Robots and Unmanned Systems, pp. 353–364. Springer (2017)

20. Lee, J., Moray, N.: Trust, control strategies and allocation of function in human-machine systems. Ergonomics **35**(10), 1243–1270 (1992)
21. Lee, J.D., Kolodge, K.: Exploring trust in self-driving vehicles through text analysis. Human factors p. 0018720819872672 (2019)
22. Lee, J.D., Moray, N.: Trust, self-confidence, and operators' adaptation to automation. International journal of human-computer studies **40**(1), 153–184 (1994)
23. Lee, J.D., See, K.A.: Trust in automation: Designing for appropriate reliance. Human factors **46**(1), 50–80 (2004)
24. Lu, Y., Sarter, N.: Eye tracking: A process-oriented method for inferring trust in automation as a function of priming and system reliability. IEEE Transactions on Human-Machine Systems (2019)
25. Metcalfe, J., Marathe, A., Haynes, B., Paul, V., Gremillion, G., Drnec, K., Atwater, C., Estepp, J., Lukos, J., Carter, E., et al.: Building a framework to manage trust in automation. In: Micro-and Nanotechnology Sensors, Systems, and Applications IX, vol. 10194, p. 101941U. International Society for Optics and Photonics (2017)
26. Meyer, J.: Effects of warning validity and proximity on responses to warnings. Human Factors **43**(4), 563–572 (2001). DOI 10.1518/001872001775870395. URL `http://journals.sagepub.com/doi/10.1518/001872001775870395`
27. Meyer, J.: Conceptual issues in the study of dynamic hazard warnings. Human Factors **46**(2), 196–204 (2004)
28. Molnar, L.J., Ryan, L.H., Pradhan, A.K., Eby, D.W., Louis, R.M.S., Zakrajsek, J.S.: Understanding trust and acceptance of automated vehicles: An exploratory simulator study of transfer of control between automated and manual driving. Transportation research part F: traffic psychology and behaviour **58**, 319–328 (2018)
29. Mueller, S.T., Piper, B.J.: The psychology experiment building language (PEBL) and PEBL test battery. Journal of neuroscience methods **222**, 250–259 (2014)
30. Muir, B.M.: Trust between humans and machines, and the design of decision aids. International journal of man-machine studies **27**(5-6), 527–539 (1987)
31. Muir, B.M., Moray, N.: Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. Ergonomics **39**(3), 429–460 (1996)
32. Pop, V.L., Shrewsbury, A., Durso, F.T.: Individual differences in the calibration of trust in automation. Human factors **57**(4), 545–556 (2015)
33. Rempel, J.K., Holmes, J.G., Zanna, M.P.: Trust in close relationships. Journal of personality and social psychology **49**(1), 95 (1985)
34. Robert, L.P., Denis, A.R., Hung, Y.T.C.: Individual swift trust and knowledge-based trust in face-to-face and virtual team members. Journal of Management Information Systems **26**(2), 241–279 (2009)
35. SAE: SAE J3016—taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. Tech. rep., SAE International, Troy, MI (2016)
36. Saeidi, H., Wagner, J.R., Wang, Y.: A mixed-initiative haptic teleoperation strategy for mobile robotic systems based on bidirectional computational trust analysis. IEEE Transactions on Robotics **33**(6), 1500–1507 (2017)
37. Saeidi, H., Wang, Y.: Incorporating trust and self-confidence analysis in the guidance and control of (semi) autonomous mobile robotic systems. IEEE Robotics and Automation Letters **4**(2), 239–246 (2018)
38. Schaefer, K.: The perception and measurement of human-robot trust. Ph.D. thesis, University of Central Florida, Orlando, FL (2013)
39. Sheridan, T.B., Vámos, T., Aida, S.: Adapting automation to man, culture and society. Automatica **19**(6), 605–612 (1983)
40. Soh, H., Xie, Y., Chen, M., Hsu, D.: Multi-task trust transfer for human–robot interaction. The International Journal of Robotics Research p. 0278364919866905 (2019)
41. Stanton, C.J., Stevens, C.J.: Don't stare at me: the impact of a humanoid robot's gaze upon trust during a cooperative human–robot visual task. International Journal of Social Robotics **9**(5), 745–753 (2017)
42. Thropp, J.E., Oron-Gilad, T., Szalma, J.L., Hancock, P.A.: Calibrating adaptable automation to individuals. IEEE Transactions on Human-Machine Systems **48**(6), 691–701 (2018). DOI 10.1109/THMS.2018.2844124
43. de Visser, E.J., Peeters, M.M., Jung, M.F., Kohn, S., Shaw, T.H., Pak, R., Neerincx, M.A.: Towards a theory of longitudinal trust calibration in human–robot teams. International Journal of Social Robotics pp. 1–20 (2019)
44. Wagner, A.R., Robinette, P., Howard, A.: Modeling the human-robot trust phenomenon: A conceptual framework based on risk. ACM Transactions on Interactive Intelligent Systems (TiiS) **8**(4), 1–24 (2018)
45. Wang, G., Wang, B., Wang, T., Nika, A., Zheng, H., Zhao, B.Y.: Defending against sybil devices in crowdsourced mapping services. In: Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services, pp. 179–191 (2016)
46. Ward, C., Raue, M., Lee, C., D'Ambrosio, L., Coughlin, J.F.: Acceptance of automated driving across generations: The role of risk and benefit perception, knowledge, and trust. In: International Conference on Human-Computer Interaction, pp. 254–266. Springer (2017)
47. Wickens, C., Dixon, S., Goh, J., Hammer, B.: Pilot dependence on imperfect diagnostic automation in simulated uav flights: An attentional visual scanning analysis (tech rep. no. ahfd-05-02). Urbana-Champaign, IL: Univ. of Illinois **21**(3), 3–12 (2005)
48. Wickens, C.D., Dixon, S.R., Johnson, N.R.: UAV automation: Influence of task priorities and automation imperfection in a difficult surveillance task. Aviation Human Factors Division, Institute of Aviation, University of Illinois at Urbana-Champaign, 2005, Chicago, IL (2005)
49. Wickens, C.D., Gordon, S.E., Liu, Y., et al.: An introduction to human factors engineering. Longman New York (1998)
50. Xu, A., Dudek, G.: Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations. In: 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 221–228. IEEE (2015)
51. Yagoda, R.E., Gillan, D.J.: You want me to trust a ROBOT? The development of a human–robot interaction trust scale. International Journal of Social Robotics **4**(3), 235–248 (2012)
52. Yang, X.J., Unhelkar, V.V., Li, K., Shah, J.A.: Evaluating effects of user experience and system transparency on trust in automation. In: 2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI, pp. 408–416. IEEE (2017)
53. Zhang, T., Tao, D., Qu, X., Zhang, X., Lin, R., Zhang, W.: The roles of initial trust and perceived risk in public's acceptance of automated vehicles. Transportation research part C: emerging technologies **98**, 207–220 (2019)
54. Zhao, H., Azevedo Sá, H., Esterwood, C., Yang, X.J., Robert, L., Tilbury, D.: Error type, risk, performance and trust: Investigating the impacts of false alarms and

misses on trust and performance. In: In Proceedings of the Ground Vehicle Systems Engineering and Technology Symposium (GVSETS 2019), pp. 1–8. NDIA, Novi, MI (2019)