Modeling the Interplay of Trust and Attention in HRI: an Autonomous Vehicle Study

Indu P Bodala  
indu@comp.nus.edu.sg  
National University of Singapore

Weicong Sng  
weicong@comp.nus.edu.sg  
National University of Singapore

Bing Cai Kok  
kokbc@comp.nus.edu.sg  
National University of Singapore

Harold Soh  
harold@comp.nus.edu.sg  
National University of Singapore

Abstract
In this work, we study and model how two factors of human cognition, trust and attention, affect the way humans interact with autonomous vehicles. We develop a probabilistic model that succinctly captures how trust and attention evolve across time to drive behavior, and present results from a human-subjects experiment where participants interacted with a simulated autonomous vehicle while engaging with a secondary task. Our main findings suggest that trust affects attention, which in turn affects the human’s decision to intervene with the autonomous vehicle.

Keywords
probabilistic models; trust; attention; autonomous vehicles

1 Introduction
Despite autonomous vehicles (AVs) rapidly gaining autonomous capabilities, human monitoring or intervention remains important to ensure safety. Understanding how human cognition affects such interactions with AVs [4, 5, 7] is crucial for the design of systems that mitigate errors and encourage proper usage.

In this late breaking report, we investigate how two major aspects of human cognition, trust and attention, influence human-AV interactions. Previous work [12] had examined experimentally how various cognitive factors moderate trust in automation. In contrast, we seek to formally model the dynamic relationship between trust and attention as the human interacts with the AV. Such computational models can be used to explain human behavior, guide AV design, or to support robot decision-making [3].

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/authors(s).

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/authors(s).

ACM Reference Format:

Figure 1: Probabilistic graphical model for the latent trust and attention model (specialized for our AV experiment). Unshaded nodes represent latent (unobserved) random variables. Observed variables are shaded. Due to model complexity, functional parameters used in the model are not shown. A decision to takeover control $O(k)$ depends on the participant’s trust $T_k$, their level of attention $A_k$, and the type of critical event $C_k$. Abbreviations for the remaining variables: $S$ - Subjective Trust Rating; $B$ - Perceived Critical Event; $E$ - Eyetrack-; $W$ - Workload condition; $Y$ - Outcome of Critical Event.

Figure 2: Map of the two simulated AV environments, annotated with critical events. Best viewed in color.

To this end, we develop a probabilistic model that succinctly captures how trust and attention evolve across time. Due to space constraints, we focus our discussion on the model’s salient features and our main findings from a human-subjects study designed to validate the model. We conduct a Bayesian analysis and find that trust affects attention ($T_{k-1} \rightarrow A_k$). In turn, both trust and attention affect the human’s decision to takeover control.

2 A Probabilistic Model of Trust and Attention
Our model aims to describe how humans interact with AVs across time. Such interactions can involve a learning phase and an interaction phase. In the former, participants passively observe the AV (e.g. observing a demonstration). In the subsequent interaction phase, participants are granted control over the vehicle, and may still learn and update their trust during this second phase. Similar to many human-robot collaboration tasks, there is a degree of risk involved; modern AVs are still prone to mistakes and participants can choose to initiate a takeover if they distrust the AV.
were allowed to intervene, that is, they could override the operations with respect to the robot [3, 14]. We conceptualize attention as with a robot/AV and captures one’s willingness to be vulnerable to future work. Comprehensive and validated trust scores [13] but defer its analysis on the y-axis indicate the priors used. (A) Credible intervals for \( \beta_{T \rightarrow A} \). (B) Credible intervals for \( \beta_{W \rightarrow A}^{\text{HL-NL}} \) and \( \beta_{W \rightarrow A}^{\text{ML-NL}} \).

Critically, the fact that such human-AV interactions unfold over a relatively protracted period of time brings the human’s state of attention into play. Our model explicitly captures this dynamic interplay between trust and attention. We adopt the definition of trust as a latent variable that summarizes the past interactions with a robot/AV and captures one’s willingness to be vulnerable with respect to the robot [3, 14]. We conceptualize attention as the spotlight of our awareness, i.e., the process by which sensory stimuli is selected for preferential processing [1]. in the context of human-AV interaction, this corresponds to how much attention one allocates to monitoring the behavior of the AV relative to some other distractor task (e.g., browsing one’s phone).

Fig. 1 illustrates a model specialized for our experiment (Sec. 2.1) where we collected eye-tracking data, subjective trust ratings, and whether the participants perceived critical events. The trust (\( T_{t-1} \rightarrow T_t \)) and attention (\( A_{t-1} \rightarrow A_t \)) updates were modelled as deterministic functions. All other relations were modelled probabilistically, e.g., workload (\( W_t \)) induced by a secondary task impacts attention, and both trust and attention affect takeovers (\( O_{\text{ck}} \)).

2.1 Validation with Human-Subjects Study

Our human-subjects study was designed to test the veracity of our model; specifically, whether the proposed model is consistent with empirical data (H1), the workload experienced by subjects affects attention (H2), and that trust in the AV affects attention (H3).

Experimental Design. Participants engaged with a simulated AV using the CARLA [6] simulator. We used 2 different town maps (Fig. 2) for the 2 different phases of the experiment. Participants were required to perform a secondary arithmetic task displayed on a second screen while engaging with the AV.

To assess their state of attention, we used a binocular Pupil Labs eye-tracking headset [11], and surface tracking to automatically label whether the subjects were looking at the AV simulator screen.

As per the model, the experiment was split into two phases. In the first phase, participants were only required to indicate if they perceived a critical event—their actions had no effect on the actual outcome of the critical event. In the subsequent phase, participants were allowed to intervene, that is, they could override the operations of the AV, thereby affecting the outcome of a critical event. After each critical event, participants were prompted to report their level of trust in the AV on a continuous scale that ranged from 0 (no trust in AV) to 100 (complete trust in AV). We also collected more comprehensive and validated trust scores [13] but defer its analysis to future work.

Study Procedure. 48 participants (Age=23.95 ± 3.16 years; 17 Female) with valid driving licenses participated in the study. We removed 3 subjects that had incomplete behavioral data, and removed 9 subjects whose eye-tracking data were unavailable due to technical issues with the eye-tracker.

For the first phase, subjects were randomly assigned to one of three experimental groups. In the No-Load (NL) condition, there was no secondary task. The secondary task was designed to be either medium-paced in the Medium-Load (ML) condition or fast-paced in the High-Load (HL) condition. All participants were then subjected to the ML condition for the second phase. Participants were awarded points for every arithmetic problem they solved in the secondary task, and penalized for missing critical events or crashes. As a manipulation check, we assessed participants’ subjective task load with NASA-TLX [9] at the end of the learning phase. The Physical Demand subscale was ignored in our analysis as our task involved minimal physical effort. The scores for the remaining subscales were summed to derive a total workload score.

3 Results

For the manipulation check, we conducted a one-way Bayesian ANOVA with the total workload score as the dependent variable in JASP v0.11.1 [10]. The results suggest that our workload manipulation had an effect (Bayes Factor = 1.915), with the data almost twice as likely under the one-way model than the null model. Post-hoc analyses with pairwise Bayesian t-tests, controlling for multiple comparisons [16, 18], suggest that the ML group experienced increased workload relative to the NL group (Adjusted Posterior Odds = 5.982), but there was little evidence for any differences in the other comparisons (Adjusted Posterior Odds < 1). For our graphical model, we inferred the posterior distribution of the latent variables using the Hamiltonian Monte Carlo algorithm in PyStan v2.19 [2, 15]. Weakly informative priors were used for all unobserved quantities [2]. We ran 6 chains with 8000 iterations and 6000 burn-in, and assessed convergence via standard tests.

Following previous work [8, 17], we generated predictions from the posterior predictive distributions to test H1. The mean posterior predictions for variables \( S \) (trust) and \( E \) (eyetracking) were highly correlated with the actual data \( r_{S}^{\text{Phase 1}} = 0.79 \), \( r_{E}^{\text{Phase 1}} = 0.89 \), \( r_{S}^{\text{Phase 2}} = 0.98 \), \( r_{E}^{\text{Phase 2}} = 0.97 \). Mean posterior predictions for the binary variables B and O yielded AUCs of 0.71 and 0.67 respectively, suggesting that our model describes the data well.

We queried the posterior distribution of our graphical model to test hypotheses H2 and H3. These two hypotheses are related to the directed edges (\( W \rightarrow A \)) and (\( T \rightarrow A \)) respectively. Let \( W_{k,j} = 1 \) if participant \( j \) was in group \( k \) and \( W_{k,j} = 0 \) otherwise. The functional form that involves these edges is \( A_{t-1,j} + \beta_{T \rightarrow A} A_{t-1,j} = \beta_{W \rightarrow A} W_{(t-1),j} + \beta_{H_{L \rightarrow NL}} H_{L_{j,t-1}} - \beta_{W \rightarrow A}^{\text{HL-NL}} W_{(t-1),j} - \beta_{W \rightarrow A}^{\text{ML-NL}} W_{(t-1),j} \). The parameters of interest are \( \beta_{T \rightarrow A} \), \( \beta_{W \rightarrow A}^{\text{HL-NL}} \) and \( \beta_{W \rightarrow A}^{\text{ML-NL}} \). Their corresponding 95% credible intervals are shown in Fig. 3. All credible intervals do not include zero, suggesting that both trust and experienced workload affect attention. Crucially, the negative sign of the coefficients imply (1) that higher trust leads to lower levels of attention and (2) the higher workload induced by the secondary task leads to lower attention. A sensitivity indicated that the results are robust to the choice of priors (Fig 3).
References


