# 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

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

Weicong Sng
weicong@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

**ACM Reference Format:**
Indu P Bodala, Bing Cai Kok, Weicong Sng, and Harold Soh. 2020. Modeling the Interplay of Trust and Attention in HRI: an Autonomous Vehicle Study. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20 Companion), March 23–26, 2020, Cambridge, United Kingdom.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3371382.3378262

## 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].

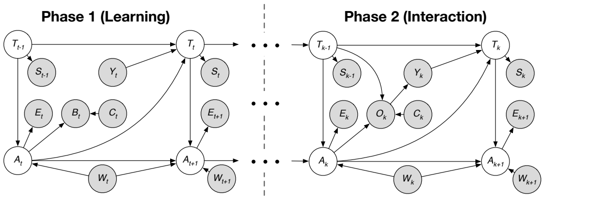

**Figure 1: Probabilistic graphical model for the latent trust and attention model (specialized for our AV experiment). Unshaded nodes represent latent (unobserved) random variables. Observed variables are shaded. Due to model complexity, functional parameters used in the model are not shown. A decision 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; $E$ - Eyetracking; $W$ - Workload condition; $Y$ - Outcome of Critical Event.**

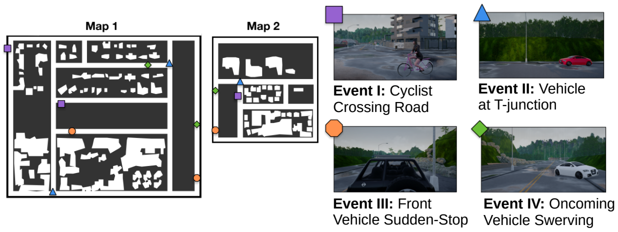
Map 1
Map 2
Event I: Cyclist Crossing Road
Event II: Vehicle at T-junction
Event III: Front Vehicle Sudden-Stop
Event IV: Oncoming Vehicle Swerving

**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 ($T_{t-1} \rightarrow A_t$). 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.

Indu P. Bodala is currently at the Department of Computer Science and Technology, University of Cambridge, UK. e-mail: ipb29@cam.ac.uk.

*HRI '20 Companion, March 23–26, 2020, Cambridge, United Kingdom*

ACM ISBN 978-1-4503-7057-8/20/03.
https://doi.org/10.1145/3371382.3378262

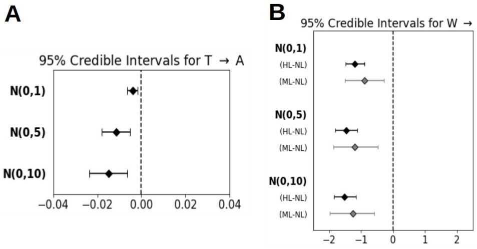

**Figure 3: 95% credible intervals for parameters of interest. Captions on the y-axis indicate the priors used. (A) Credible intervals for $\beta_{T \to A}$. (B) Credible intervals for $\beta^{HL-NL}_{W \to A}$ and $\beta^{ML-NL}_{W \to A}$.**

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 ($T_{t-1} \longrightarrow T_t$) and attention ($A_{t-1} \longrightarrow A_t$) 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 ($O_k$).

## 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 ($r_S^{Phase\ 1} = 0.79$, $r_S^{Phase\ 2} = 0.89$, $r_E^{Phase\ 1} = 0.98$, $r_E^{Phase\ 2} = 0.97$). 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 ($W \longrightarrow A$) and ($T \longrightarrow A$) 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 $A_{t,j} = A_{(t-1),j} + \beta_{T \to A} T_{(t-1),j} - \beta^{HL-NL}_{W \to A} \big( W^{(t)}_{HL,j} - W^{(t-1)}_{HL,j} \big)$ where the parameters of interest are $\beta_{T \to A}$, $\beta^{HL-NL}_{W \to A}$ and $\beta^{ML-NL}_{W \to A}$. 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

[1] Christopher L Asplund, J Jay Todd, Andy P Snyder, and René Marois. 2010. A central role for the lateral prefrontal cortex in goal-directed and stimulus-driven attention. *Nature neuroscience* 13, 4 (2010), 507.

[2] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. Stan: A probabilistic programming language. *Journal of statistical software* 76, 1 (2017).

[3] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. 2018. Planning with trust for human-robot collaboration. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 307–315.

[4] ML Cummings and AS Clare. 2015. Holistic modelling for human-autonomous system interaction. *Theoretical Issues in Ergonomics Science* 16, 3 (2015), 214–231.

[5] Joost CF De Winter, Riender Happee, Marieke H Martens, and Neville A Stanton. 2014. Effects of adaptive cruise control and highly automated driving on workload and situation awareness: A review of the empirical evidence. *Transportation research part F: traffic psychology and behaviour* 27 (2014), 196–217.

[6] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An open urban driving simulator. *arXiv preprint arXiv:1711.03938* (2017).

[7] Donald L Fisher, Maura Lohrenz, David Moore, Eric D Nadler, and John K Pollard. 2016. Humans and intelligent vehicles: the hope, the help, and the harm. *IEEE Transactions on Intelligent Vehicles* 1, 1 (2016), 56–67.

[8] Noah D Goodman, Joshua B. Tenenbaum, and The ProbMods Contributors. 2016. Probabilistic Models of Cognition. http://probmods.org/v2. Accessed: 2019-12-9.

[9] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[10] JASP Team. 2019. JASP (Version )[Computer software]. https://jasp-stats.org/

[11] Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction. In *Adjunct Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14 Adjunct)*. ACM, New York, NY, USA, 1151–1160. https://doi.org/10.1145/2638728.2641695

[12] Luke Petersen, Lionel Robert, Jessie Yang, and Dawn Tilbury. 2019. Situational awareness, driver's trust in automated driving systems and secondary task performance. *SAE International Journal of Connected and Autonomous Vehicles, Forthcoming* (2019).

[13] Kristin E Schaefer. 2016. Measuring trust in human robot interactions: Development of the "trust perception scale-HRI". In *Robust Intelligence and Trust in Autonomous Systems*. Springer, 191–218.

[14] Harold Soh, Yaqi Xie, Min Chen, and David Hsu. 2019. Multi-task trust transfer for human–robot interaction. *The International Journal of Robotics Research* (2019). https://doi.org/10.1177/0278364919866905

[15] Stan Development Team et al. 2018. PyStan: the Python interface to Stan. (2018).

[16] Don van den Bergh, Johnny van Doorn, Maarten Marsman, Tim Draws, Erik-Jan van Kesteren, Koen Derks, Fabian Dablander, Quentin F Gronau, Šimon Kucharský, Akash Raj, and et al. 2019. A Tutorial on Conducting and Interpreting a Bayesian ANOVA in JASP. https://doi.org/10.31234/osf.io/spreb

[17] Natalia Vélez and Hyowon Gweon. 2019. Integrating incomplete information with imperfect advice. *Topics in cognitive science* 11, 2 (2019), 299–315.

[18] Peter H Westfall. 1997. Multiple testing of general contrasts using logical constraints and correlations. *J. Amer. Statist. Assoc.* 92, 437 (1997), 299–306.