The Impact of Trust in Automation on Safety Outcomes: A Study on Takeover Scenarios in Conditional Driving Automation Simulator Experiments

Camila Correa-Jullian[a], Xu Han[b], Ali Mosleh[a], Jiaqi Ma[b]
[a]The B. John Garrick Institute for the Risk Sciences, University of California Los Angeles, United States, ccorreaj@ucla.edu
[b] Mobility Lab, University of California Los Angeles, United States

Abstract: Advancing our understanding of Human-Autonomy Teams (HATs) is critical for evaluating the safety of Automated Driving Systems (ADS) in real-world operations. Drawing from Human Reliability Analysis (HRA) discipline, Performance Shaping Factors (PSFs) such as trust, attention, and task load influence how human drivers and automated agents jointly perform Dynamic Driving Tasks (DDTs) during control transitions. This paper focuses on the longitudinal evolution of trust in automation across driving simulator experiments focusing on time-constrained takeover scenarios and assesses the impact of warning availability and traffic congestion on perceived risk. Mid-fidelity driving simulator-based experiments collected subjective and objective data on driver-ADS teaming during takeover scenarios involving silent automation failures characterized by tailgating behavior, with and without warnings, under low- and high-traffic conditions. Questionnaires were administered before, during, and after exposure to the simulated scenarios to capture driver profile characteristics, in-task trust perception, and longitudinal changes in trust in driving automation. Longitudinal trust (12-item adapted Trust in Automated Driving scale) was analyzed using Wilcoxon Signed-Rank tests with effect size estimation and Cumulative Link Mixed Models (CLMM), while within-session perceived trust (3-item scale) was administered after each scenario and analyzed using Linear Mixed-Effect Models (LMMs) and Repeated Measures ANOVA (RM-ANOVA). Results indicate statistically significant decreases in trust following exposure to automation failure behavior, particularly regarding contextual complexity and system reliability, with warning availability emerging as the dominant PSF shaping in-task perceived trust. These findings highlight the dynamic nature of trust in driver-ADS HATs and support the integration of trust-related PSFs into probabilistic risk assessment frameworks for automated driving systems.

1. INTRODUCTION

Despite the expected growth of automated vehicle deployment, human involvement will remain essential, whether as active drivers, on-board passengers, or remote fleet supervisors. Ongoing research in human-vehicle interaction has increasingly examined the safety implications of Automated Driving Systems (ADS) (SAE Levels 3-5) [1], such as passenger cars or in commercial freight vehicles. These applications require robust collaborative driver-system design approaches, built on sophisticated information sharing and decision-making strategies, as well as risk-aware task allocation and switching mechanisms appropriate for shared-autonomy settings, even under nominal operational conditions [2], [3]. In this context, Human-Autonomy Teams (HAT) [4] emerge as a promising avenue to study human-autonomous system interactions through the lens of Human Reliability Analysis (HRA) methods, offering a path to quantify the impacts that ADS design has on operational safety. Recent research into human-system interaction has increasingly emphasized collaborative and team-oriented dynamics, such as task division and allocation strategies, human's trust in automation, attention management, and the challenges associated with the explainability of the autonomous systems' decisions [5]–[7]. This is particularly relevant in the case of automation failure, where factors such as system design and limited explainability of automated decision-making may hinder the capacity of the human operator or user to interpret and respond accordingly [8], [9]. In this regard, the user population matters considerable: on-

board drivers may range from trained safety drivers to inexperienced everyday users, with the latter being more exposed to trust calibration issues and mismatched expectations regarding system behavior [10], [11]. Other key aspects of human-automation relationships have been studied extensively, such as how trust (mistrust or over-trust) impacts human performance and system safety over time [12], [13]. Several studies have explored trust calibration, conflicts of control authority, regaining and maintaining situation awareness, take-over control mechanisms and the design of human-system interface (HSI) [14]–[16]. These studies, generally conducted in driving simulator environments for partial (L2) and conditional (L3) automated driving, employ a variety of methods to measure human performance, including recording reaction time to unexpected scenarios, using physiological signals to assess driver state, and debriefing questionnaires to characterize the driver perception [17]. Trust is also recognized as a fundamental element of team cohesion in the HAT context, encompassing the team's diversity, mutual trust, team training and experience, shared goals, authority gradients, and compliance with procedures [18]. For drivers and remote operators, this relates to personal experience (added to the influence of training, if applicable), the belief in shared goals (i.e., safety), comfort interacting with and trust towards the autonomous agent [19], [20].

This work presents selected findings from driver-in-the-loop simulation study examining how drivers interact with a Level 3 ADS vehicle that experiences an automation failure, represented by an onset of tailgating behavior by the ADS [21]. Experiments were implemented in the OpenCDA open-source simulation platform, focusing on capturing both objective and subjective measures quantifying driver-ADS team interactions, errors, and failures. The scenarios are designed such that researchers can assess the driver's ability to detect issues (i.e., tailgating behavior), react appropriately (e.g., decide to intervene in the vehicle's operation), and resume manual control to avoid a collision, based on two conditions: the availability of visual warnings (Factor A: Takeover Request/Warning Availability) and the complexity of the driving environment (Factor B: Traffic Level). A repeated-measures design with counterbalancing was used, exposing each participant to all four scenario conditions in randomized order. This paper focuses specifically on the longitudinal dynamics of trust in automation, drawing on questionnaire data collected before, during, and after simulator exposure.

2. EXPERIMENTAL DESIGN

This section provides an overview of the experiment protocol, data collection methods, and PSF model quantification efforts. Experiments focused on how participants transition between manual and automated driving modes when experiencing automation failure, represented by an onset of tailgating behavior by the ADS. The simulated driving scenarios were designed to assess the driver's ability to detect issues, react appropriately, and resume manual control effectively. The specific objective of the experiments is to assess the effect of scenario complexity ("External Environment") and the availability of early warnings ("Conditioning Events") over driver performance in takeover scenarios, i.e., when the driver is expected to switch from automated to manual driving. Questionnaires are used to extract the participants' driver profile, trust in driving automation, and perception of the simulated driving scenarios. These subjective ratings can be combined with simulation data, including vehicle kinematics, reaction times, and recorded incidents, to derive the impact of agent, scenario, and team PSFs on safety performance.

2.1 Experiment Setup

The experimental setup consists of a medium-fidelity driving simulator based on OpenCDA's virtual simulation platform, and physical controls and hardware, including a Logitech G923 steering wheel and pedal bundle, a Next Level Racing F-GT simulator cockpit and seat, and a triple 32-inch monitor configuration (**Figure 1**). OpenCDA is a comprehensive open-source ecosystem that includes a collection of pre-trained models, a range of simulators for driving and traffic at different levels of detail, benchmark datasets for training and testing, and a scenario database and generator for Cooperative Driving Automation (CDA) research [22], [23]. To conduct the driving simulator study, a new module dedicated to human-system interaction modeling was developed based on the existing functionalities of both OpenCDA and CARLA platforms.

Figure 1: The driving simulator setup and the virtual driving environment OpenCDA

2.2 Experiment Protocol

Participants were required to attend an in-person appointment, where they met with the research team, reviewed the study information sheet, and had the opportunity to ask clarifying questions. The experimental procedure had an approximate duration of 90 minutes and consisted of the following main stages: (1) an initial questionnaire to collect driving experience, habits, and opinions on automated driving technology, (2) a training session to introduce the simulator and purpose of the study, (3) a test session consisting of four scenarios with a test survey administered after every run, and (4) a final questionnaire leading to the end of the study, repeating a section of the initial survey to enable pre-post comparison of trust in driving automation. Participants were thanked and provided with compensation ($15 Amazon gift card). If any sign of motion sickness was expressed, trials were terminated early. The sessions are described as follows:

Simulator training: Participants received instructions on how to operate the driving simulator and the characteristics of the simulated vehicle's capabilities. The experiment stages were briefly explained, including a description of the Levels of Driving Automation, focusing on the differences between Levels 2, 3, and 4. Participants went through a short training session to familiarize themselves with the simulator's visuals and controls. The monitors showed a virtual driving environment representing a small town with residential and highway sections (CARLA Town 06), with traffic signs and other vehicles following independent trajectories. The monitors presented an approximate view from the driver's seat, indicating the automation state (i.e., OpenCDA or Manual) and the vehicle's speedometer **(Figure 2**). All scenarios were initiated in automated driving mode, and participants were instructed to brake or use the steering wheel to take control when desired. If any collision occurred during the training scenarios, the researcher manually terminated the simulation. This training session was divided into three short tasks:

- **Training scenario 1 - Manual control**: Participants are tasked to take manual control by braking or steering and then drive as they would in a typical vehicle down a road, interacting with other vehicles and traffic signs along three intersections.
- **Training scenario 2 - Observe automated mode:** The participants are tasked to watch how the vehicle performed the same route while maintaining a minimum safety distance of $TTC = 7$ seconds. Participants were encouraged to monitor the vehicle's actions to familiarize themselves with its expected behavior.
- **Training scenario 3 - Respond to a takeover request**: Participants are tasked to respond to a takeover request triggered at an unknown point in the track when $TTC < 5$. This visual alert appeared on the central monitor with the following text: "Warning: Please Brake" (Figure 3a). Participants are instructed to bring the vehicle to a safe stop. If a collision occurred, a visual alert appeared with the following text: "Warning: Collision Occurred" (Figure 3b) and the simulation was terminated.

Testing scenarios: Participants were then given instructions to complete short driving scenarios where they could switch between manual and automated driving controls. Each trial began with the vehicle

driving in automated mode; participants were permitted to take control at any time, although automated mode was preferred. Each trial assessed two independent variables: Factor A warning availability (A0 available, A1 not available) and Factor B traffic complexity (B0 low traffic, B1 high traffic). The four trials each covered a unique combination of these factors (A0B0, A1B0, A0B1, A1B1), presented in counterbalanced randomized order. Each trial had a duration of approximately 5 minutes. The ego vehicle followed a lead vehicle maintaining a safe distance at approximately 35-38 mph (56-61 km/h), defined as $TTC > 7$. At a random time between minutes (2-4), the safety distance was artificially reduced to 3.5 seconds, such that when the lead vehicle slowed (e.g., at an intersection), it invaded the safety envelope of the ego vehicle. When the TTC dropped below 5 seconds, this triggered (A0) or did not (A1) an a visual Takeover Request (TOR). Participants were expected to slow or stop the vehicle in a safe spot. If no takeover occurred, the ego vehicle attempted to maintain the 3.5-second distance. Scenarios ended when the driver collided with an object after regaining control or successfully completing the track. After each scenario, participants completed a 5-minute between-scenario survey before the simulation was reset and the next trial commenced. This process is repeated 4 times.

Figure 2: Experimental Setup - View from the driver's seat

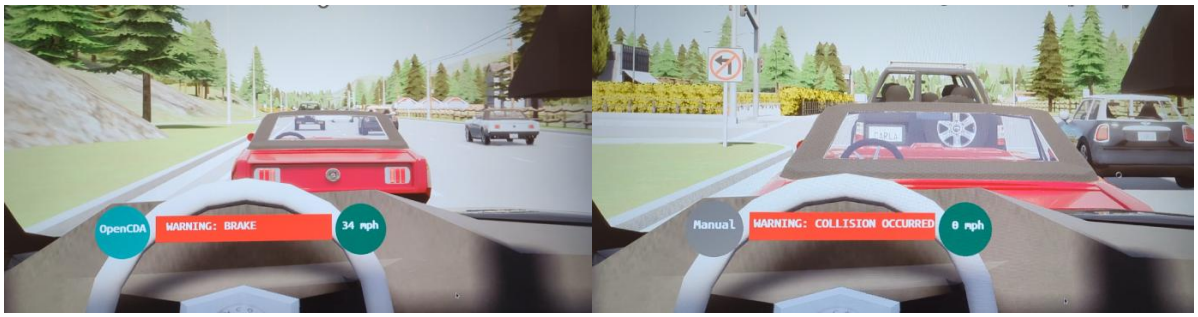

Figure 3: Example of driver interface, (a) warning alert and (b) collision alert

2.3 Data Collection

Two distinct questionnaire instruments were used to capture trust-related data at different time scales: (1) a longitudinal pre/post instrument assessing overall trust in driving automation before and after the full experiment, and (2) a brief within-session instrument administered after each of the four test scenarios to track in-task trust perceptions, among other questions. The initial and final survey included a 12-item trust scale adapted from the Trust in Automated Driving (TiAD) Questionnaire [18], [24], [25], employing a 7-point Likert response format (**Table 1**). These items assess multiple dimensions of trust, including perceived safety, delegation willingness, and manual control preference. The between-scenario survey captured in-task perceptions following each of the four test scenarios. This survey included multiple-choice and verbal open-ended questions. The multiple-choice section covered the participant's perception of the ADS's behavior, workload, and trustworthiness. The three items specific to perceived trust are presented in **Table 2** [32]. Responses were averaged across the three items to form the Perceived Trust composite PSF, which serves as a continuous, in-task trust indicator during the experiments.

Table 1: Trust in Automation Items in Initial (1) and Final (3) Survey

Item	Variable	Question*
1.1-3.1	Trust Safe	I would feel safe in an automated vehicle.
1.2-3.2	Trust ADS	The automated driving system provides me with more safety compared to manual driving.
1.3-3.3	Trust Manual	I would rather keep manual control of my vehicle than delegate it to the automated driving system on every occasion.
1.4-3.4	Trust ADS Decision	I would trust the automated driving system decisions.
1.5-3.5	Trust ADS Capacity	I would trust the automated driving system's capacities to manage complex driving situations.
1.6-3.6	Trust Weather	If the weather conditions were bad (e.g., fog, glare, rain), I would delegate the driving task to the automated driving system.
1.7-3.7	Trust Attention	Rather than monitoring the driving environment, I could focus on other activities confidently.
1.8-3.8	Trust Boring	If driving were boring for me, I would rather delegate it to the automated driving system than do it myself.
1.9-3.9	Trust Tired	I would delegate the driving to the automated driving system if I was tired.
1.10-3.10	Trust Passengers	If I had passengers in my automated car, I would rather drive by myself than delegate it to the automated driving system.
1.11-3.11	Trust TOR	I would like to recover control from the automated pilot if I did not like the way it drives.
1.12-3.12	Trust TOQ	Globally, I trust my capacity to resume control if needed.

* Scale Average [1-7]; How much do you agree with the following statements? Items 1.1–3.1 through 1.12–3.12 were administered in the initial (1) and final (3) surveys.

Table 2: Within-Session Perceived Trust Items

Item	Variable	Question
2.1	Understand ADS	To what extent do you think... you could understand the system?
2.2	Agree with ADS	To what extent do you think... you agree with the system's decision?
2.3	Trust the system	To what extent do you think... the system is trustworthy?

2.4 Data Analysis

This section details the statistical methods applied to the trust-related data collected during the experiment. Two analytical streams are reported: (1) within-session perceived trust, analyzed using LMMs and RM-ANOVA, and (2) longitudinal trust change, analyzed using the Wilcoxon Signed-Rank test and CLMM. Linear Mixed-Effects Models (LMMs) were used to analyze scenario-level PSFs constructed from aggregated Likert-type data (i.e., Perceived Trust). LMMs include both fixed effects (warning availability, traffic complexity, and their interaction) and random effects (subject-level intercepts), accounting for repeated observations per participant. Normality of residuals was assessed using the Shapiro-Wilk test, and homoscedasticity was evaluated with the Breusch-Pagan test. In addition, RM-ANOVA is used to evaluate main and interaction effects across the 2x2 within-subject design for constructed PSFs. Partial eta-squared is reported as effect size and post-hoc comparisons were conducted using Bonferroni correction. Regarding the longitudinal trust study, Wilcoxon Signed-Rank test and CLMMs are employed to test the significance of time as a contributing factor to the difference between initial and final answers. Both tests are well suited for Likert-scale responses, which are ordinal and often not normally distributed.

3. RESULTS

A total of 72 volunteers participated in the research study conducted during a 3-month period. Each participant experienced the four treatments, and after each experiment, the validity of the session was evaluated based on vehicle trajectory, warning triggers, and simulation stability. Abnormal scenarios were omitted, e.g., OpenCDA simulation was interrupted, or re-engaging the ADS resulted in failure

modes out of the scope of the experiment. After filtering out invalid experiments, the participant pool was reduced to 50 entries, ensuring each treatment order was represented by at least two participants. The sample of 50 contained 32 male (64%) and 17 female (34%) participants. Ages ranged from 19 to 39 years ($M = 25.68, SD = 5.14$) and driving experience since license acquisition ranged from 2 to 21 years ($M = 7.72, SD = 4.82$). The participant pool was mostly divided between using their private vehicle (46.8%) and public transportation (38.7%) as their main mode of transport. The majority reported using a private vehicle multiple times a week (54%), followed by once a week to once a month (36%). In addition, 76% of participants reported familiarity with at least one ADAS feature, with blind-spot warning (17%), cruise control (16.8%), lane-departure warnings (15.6%), and automated emergency brake systems (13.8%) being the most frequently mentioned features. Lane-keeping and lane-change assistance were mentioned 11.4% and 6.0% of the time, while adaptive cruise control was mentioned 11.4% of the time and driver monitoring systems, 6.0%. While these items were collected, alongside dedicated questions regarding the participant's driver profile, they are not included in the present analysis and are only presented for contextual purposes.

Baseline Trust Profile: The initial survey characterized participants' trust in automated driving prior to any simulator exposure. **Figure 4** shows the distribution of responses on a 7-point Likert scale ('Strongly Disagree' to 'Strongly Agree'). Items indicating preconceived trust in ADS (1.1-1.2 and 1.4-1.9) are summarized as follows: there is a moderate to high baseline trust that ADS are safe (42% Agree, 18% Slightly Agree, 14% Strongly Agree), but less indication that it is safer than manual driving (24% Agree, 20% Slightly Agree, 14% Strongly Agree). There is a general tendency to trust the ADS decision-making process (64% Agree or Slightly Agree) and participants were willing to delegate control under certain conditions for convenience: 40% Agree/30% Slightly Agree if driving was boring; 50% Agree, 22% Slightly Agree if driver was tired. However, participants were more conditional about their trust in complex driving situations (44% Agree/Slightly Disagree, 44% Disagree/Slightly Disagree) or adverse weather conditions (40% Agree, 32% Disagree). Items indicating potential mistrust or preference for manual control (1.3 and 1.10-1.12) reveal a strong preference for retaining authority to intervene (52% Strongly Agree, 46% Agree/Slightly Agree). Participants were divided regarding manual control when driving with passengers (50% prefer manual control on every occasion, 54% if there were other passengers on board) but strongly trusted their capacity to regain control effectively if required (38% Strongly Agree, 28% Agree, 24% Slightly Agree). In general, participants exhibited a conditional view of the ADS's capacities, potentially affecting their takeover performance in failure scenarios.

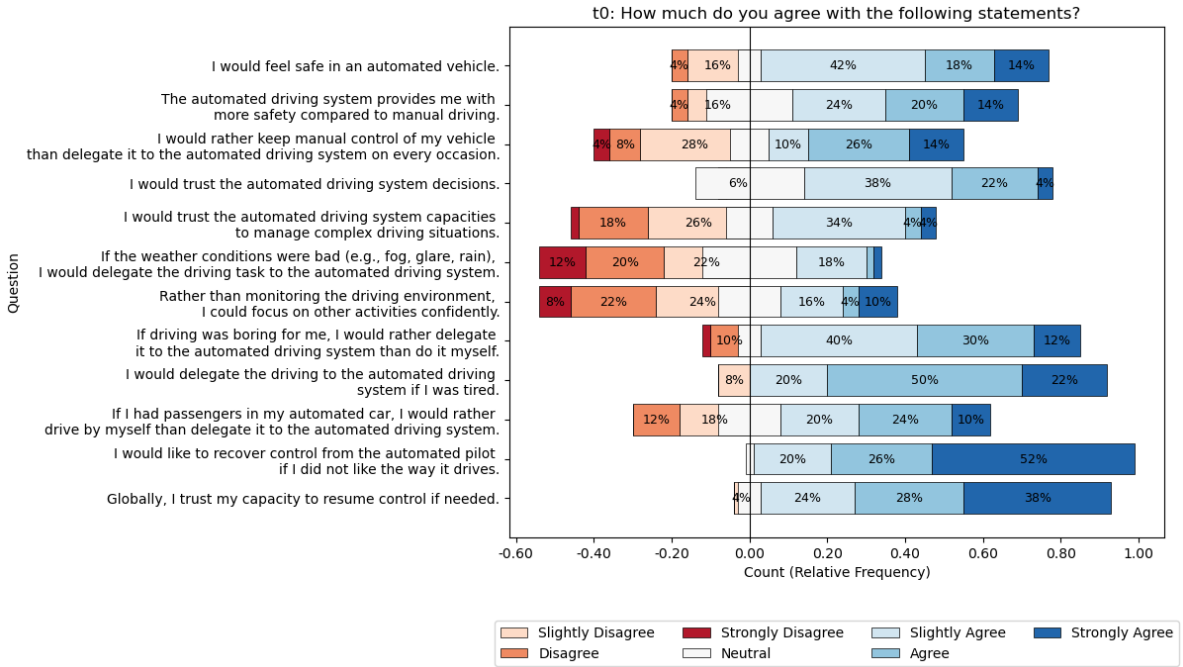

Figure 4: Initial Survey - Trust in Driving Automation

Within-Session Perceived Trust: The mid-session questionnaire captured participants' in-task perceived trust following each of the four scenario conditions: A (warning, low traffic), B (no warning, low traffic), C (warning, high traffic), and D (no warning, high traffic). Responses across the three items were averaged to form the Perceived Trust composite score (**Table 3**). Perceived Trust was highest under Condition A (M = 4.96, SD = 1.23) and lowest under Conditions B and D (M = 3.89 for both), with Condition C yielding an intermediate value (M = 4.35, SD = 1.10). This pattern suggests that warning availability is the primary driver of in-task perceived trust, while traffic complexity exerts a secondary, moderating influence.

Table 3: Descriptive Statistics for Perceived Trust by Scenario Condition (N = 50)

Condition	Max	Min	Mean	Std	Q1 (25%)	Q2 (50%)	Q3 (75%)
A – Warning, Low Traffic	7.00	2.00	4.96	1.23	4.08	5.00	6.00
B – No Warning, Low Traffic	6.33	1.67	3.89	1.19	3.00	3.84	4.67
C – Warning, High Traffic	6.67	1.67	4.35	1.10	3.67	4.33	5.00
D – No Warning, High Traffic	7.00	1.33	3.89	1.15	3.33	3.67	4.33

To assess the independent and joint effects of Factor A (warning availability) and Factor B (traffic level), an LMM was fitted to the Perceived Trust scores (**Table 4)**. The model revealed significant main effects of both warning absence (t = −5.87) and high traffic (t = −3.34), as well as a significant interaction (t = 2.36), indicating that the effect of warning availability on perceived trust is partly attenuated under high-traffic conditions. Model residuals did not deviate significantly from normality (Shapiro-Wilk p = 0.705) and showed no evidence of heteroscedasticity (Breusch-Pagan p = 0.86).

Table 4: LMM Results for Perceived Trust (MeanScore ~ FactorA * FactorB + (1 | Subject))

Term	Estimate	Std. Error	t value	Interpretation
Intercept	4.9598	0.17	30.04	Baseline (Condition A: Warning, Low Traffic)
Factor A	−1.0662	0.18	−5.87	Warning absence significantly reduces perceived trust
Factor B	−0.6058	0.18	−3.34	High traffic significantly reduces perceived trust
Interaction A×B	0.6056	0.26	2.36	Partial compensation: warning absence effect attenuated under high traffic

RM-ANOVA corroborated these findings (**Table 5**), confirming significant main effects for Factor A ($p = 2.1\times10^{-7}$, $\eta^2 = 0.375$) and Factor B ($p = 1.2\times10^{-2}$, $\eta^2 = 0.086$), as well as a significant interaction ($p = 3.7\times10^{-2}$, $\eta^2 = 0.086$). The large effect size for warning availability (0.375) confirms that warning presence is the dominant factor shaping in-task perceived trust, consistent with the LMM results. Pairwise comparisons further indicate that warning conditions yielded a mean perceived trust of 4.66 compared to 3.90 under no-warning conditions, and that low-traffic conditions yielded 4.43 compared to 4.12 under high-traffic conditions.

Table 5: RM-ANOVA Results and Model Diagnostics for Perceived Trust

Metric	Factor A: Warning	Factor B: Traffic	Interaction A×B
p-value (RM-ANOVA)	2.1×10^{-7}	1.2×10^{-2}	3.7×10^{-2}
Partial η^2	0.375	0.086	0.086
Condition means (A0/A1; B0/B1)	4.66 (warning) vs. 3.90 (no warning)	4.43 (low) vs. 4.12 (high)	Warning is dominant; interaction significant
Shapiro-Wilk p-value (residuals)	0.705	—	Residuals do not deviate from normality

Longitudinal Trust Evaluation: The final survey assessed changes in participants' trust following exposure to the full test session. **Figure 5** shows the updated distribution of responses using the same 7-point Likert scale. Compared to baseline responses, the post-test results suggest declines in trust across most ADS-related statements. Notable changes include significantly lower trust that ADS are safe (agreement declined to 28% from 74%) and that ADS provides more safety than manual driving

(agreement declined to 24% from 58%), with strong disagreement rising to 20%, indicating a marked reduction in perceived safety benefit. Participants also indicated a moderately stronger preference to maintain manual control over the ADS under all circumstances (66% total agreement from 50%). Willingness to delegate driving under specific conditions also weakened. When tired, bored, or when passengers were on board, fewer participants strongly agreed to let the ADS take over (e.g., only 6% "Strongly Agreed" when tired, down from 22%), and disagreeing responses became more frequent (e.g., 26%, up from 8%). Delegation in adverse weather showed a clear decline: "Disagree/Strongly Disagree" responses increased to 38% and 32%, respectively, while agreement dropped to around 10%. Confidence in the ability to resume control remained relatively high, though slightly tempered: "Strongly Agree" dropped from 38% to 28%, and neutral responses became more common, suggesting participants also assessed that scenario complexity would negatively affect their own performance.

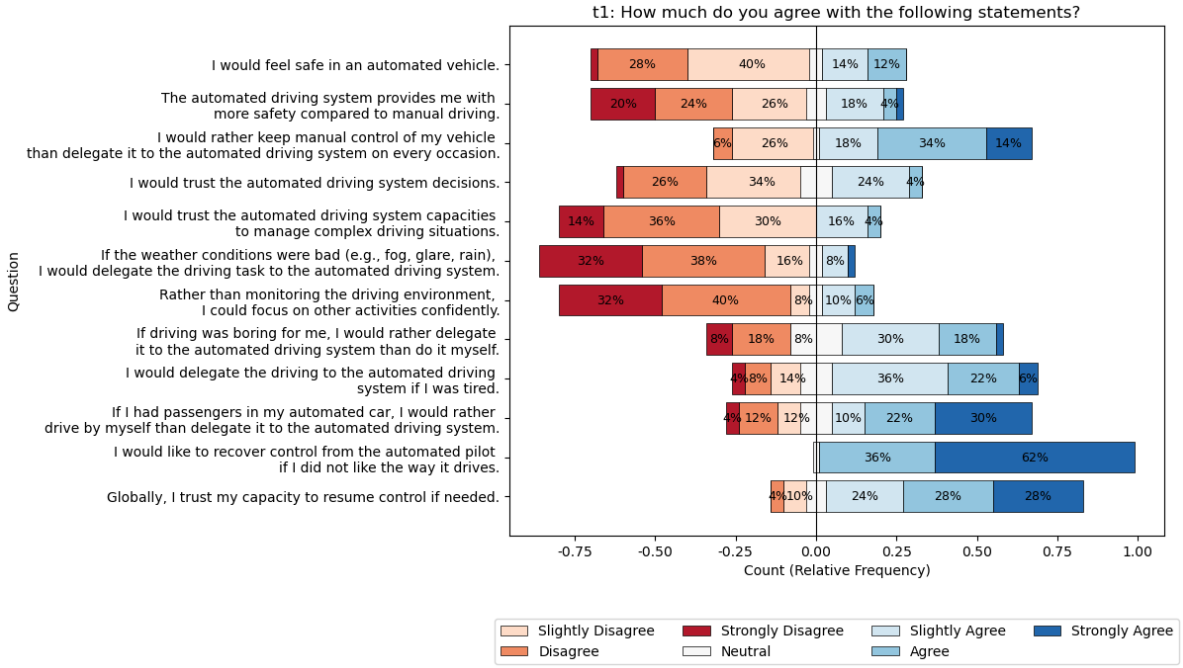

Figure 5: Final Survey - Trust in Driving Automation

Figure 6 presents a Kernel Density Estimate (KDE) plot comparing both response distributions. Across 12 trust-related Likert items, the Wilcoxon Signed-Rank test showed statistically significant changes in 10 out of 12 questions (
Table 6).

Table 6: Longitudinal Trust Test Results

Questions	Variable	W	p-value	Effect Size (r)	Interpretation
1.1-3.1	Trust Safe	65	6.24E-07	-0.705	Significant large decrease
1.2-3.2	Trust ADS	26	1.80E-07	-0.738	
1.3-3.3	Trust Manual	294.5	9.25E-02	0.000	Not significant
1.4-3.4	Trust ADS Decision	39	2.66E-07	-0.728	Significant large decrease
1.5-3.5	Trust ADS Capacity	66	6.22E-05	-0.566	
1.6-3.6	Trust Weather	67	2.22E-05	-0.600	
1.7-3.7	Trust Attention	110	7.79E-05	-0.559	
1.8-3.8	Trust Boring	116	5.88E-04	-0.486	Significant moderate-to-large decrease
1.9-3.9	Trust Tired	61.5	1.61E-05	-0.610	Significant large decrease
1.10-3.10	Trust Passengers	511	1.68E-01	0.195	Not significant
1.11-3.11	Trust TOR	168	1.31E-02	0.000	Significant but small/neutral change
1.12-3.12	Trust TOQ	107	4.59E-02	0.000	

The strongest effects were observed in: *Trust ADS* ($p < 0.0001, r = -0.738$) and *Trust ADS Decision* ($p < 0.00001, r = -0.728$). *Trust ADS Capacity, Trust Safe, Trust Attention* also showed large and significant decreases ($-0.71 \leq r \leq -0.56$). Trust under fatigue and boredom (*Trust Tired, Trust Boring*) demonstrated moderate-to-large decreases ($r = -0.610$ and $r = -0.486$, respectively). *Trust Manual* and *Trust Passengers* showed no statistically significant changes ($p > 0.05$), suggesting that perceptions of manual control and the presence of passengers were more stable across the intervention. *Trust TOQ* and *Trust TOR* reached statistical significance ($p < 0.05$), but their effect sizes were negligible ($r \approx 0$), indicating minimal practical change.

In addition, CLMM was used to evaluate trust rating evolution across time (Error! Not a valid bookmark self-reference.). The fixed effect for time ($timePost = -1.051$) indicates a significant decrease in trust ratings following the intervention, even when accounting for differences in individual subjects and question types. The substantial random variance across questions ($SD = 1.345$) suggests that not all items shifted equally, aligning with the observation that some trust items were more stable than others.

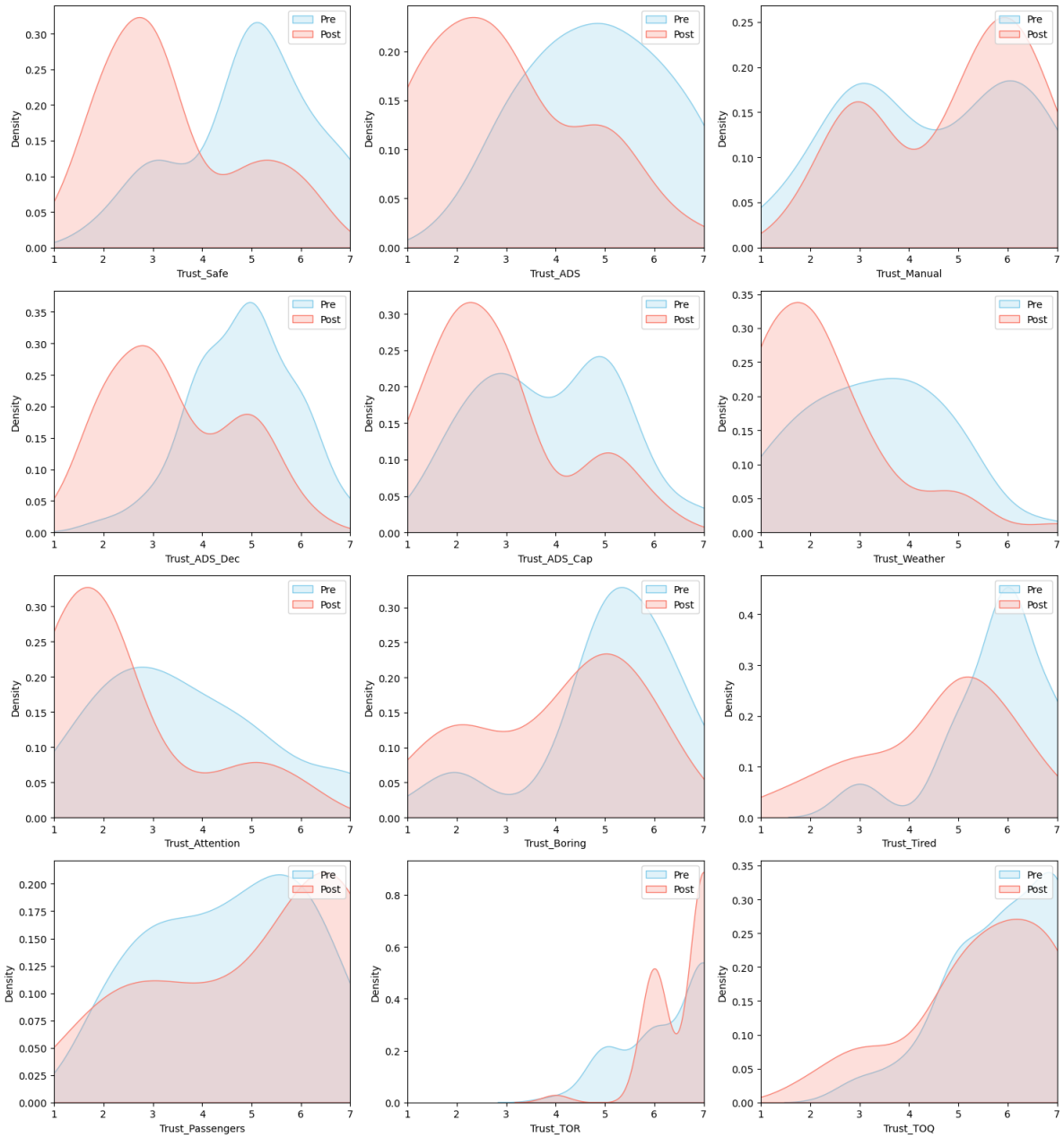

Figure 6: Kernel Density Estimate Pre-Post Trust

Table 7: Longitudinal Trust CLMM Results

Component	Variable	Estimate	Interpretation
Fixed Effect	timePost	−1.051	Significant decrease in trust ratings post-intervention (lower odds of higher responses)
Random Effects	Subject (SD)	0.5602	Moderate variability across subjects in baseline trust
	Question (SD)	1.3495	Large variability across trust questions
	Subject Variance	0.3138	Moderate variability across individuals in baseline trust levels
	Question Variance	1.821	Substantial variability across questions.
Model Fit	Log-likelihood	−1990.74	Indicates model fit quality
	AIC	3999.49	Penalized model fit; lower is better
	N Observations	1200	50 subjects × 12 questions × 2 timepoints
	Convergence (max. gradient)	5.16e−04	Indicates successful model convergence

4. DISCUSSION

As seen in similar studies, trust decreased significantly following exposure to automation failure scenarios [20], however, not all trust dimensions are equally sensitive to failure exposure. Items reflecting general preferences for manual control (*Trust Manual*) and passenger scenarios (*Trust Passengers*) showed no significant changes. Conversely, items directly tied to ADS capability, safety, and delegation—including weather, attention, and fatigue contexts—were most responsive to the intervention. This heterogeneity has direct implications for PSF model design: trust should be disaggregated where possible, separating intrinsic from extrinsic sources of trust. Within-session perceived trust (the Perceived Trust PSF) tracked the experimental conditions in a meaningful and interpretable way. Warning availability (Factor A) was the dominant driver ($\eta^2 = 0.375$, $p = 2.1\times10^{-7}$), with a significant interaction (A×B, $p = 0.037$). The pattern across conditions (A: 4.96 > C: 4.35 > B ≈ D: 3.89) suggests that the presence of a warning partially compensates for the negative effect of increased traffic complexity on in-task trust. Silent failures under high-traffic conditions (Condition D) produced the lowest perceived trust ratings and the broadest spread. Silent failures are particularly damaging to trust calibration, especially under conditions of high environmental complexity, where drivers may already face elevated cognitive demands. These results suggest that transparent, timely system communication is critical not only for safety performance but for maintaining appropriate trust levels throughout an operation. As humans interact with automated systems over time, adaptation is likely, not just to specific tasks, but to the team dynamic itself. In this study, each test session comprised multiple short scenarios (3-4 minutes each), allowing for potential post-hoc analysis of how drivers adapt over repeated interactions. While longer simulation studies have demonstrated behavioral changes over time [26]–[28], this experiment sought to capture similar dynamics within a shorter, counterbalanced format. Although scenario order and exposure time were formally controlled through randomization and training protocols, the existing dataset still allows for meaningful post-hoc analysis of time-related effects and construction of indicators, such as the number of takeovers per participant as a proxy for evolving trust or driver strategy. It should be noted that the use of a medium-fidelity simulator and convenience sample limit direct generalizability to real-world ADS interactions. Individual driver profile characteristics (e.g., ADS experience, technology readiness) were collected but not modeled in this analysis. Future work will integrate these subjective trust measures with objective simulation data and performance measures, including vehicle kinematics, reaction times, crash rates, and frequency of hard braking events, to construct and validate the complete BBN-based HAT model representing how agent- and scenario-level PSFs jointly impact safety outcomes [29].

5. CONCLUSION

This study examined trust dynamics in driver-ADS HATs using two complementary instruments: a 12-item pre/post trust scale (TiAD-adapted) and a 3-item within-session Perceived Trust measure administered after each of four experimental conditions varying warning availability and traffic complexity. Significant and practically meaningful trust reduction was confirmed across 10 of 12 items

(Wilcoxon Signed-Rank, large effect sizes) and validated by CLMM, particularly for ADS safety, decision-making, and delegation under complex or adverse conditions. Within-session perceived trust was primarily sensitive to warning availability, with a significant interaction effect further highlighting the importance of system communication under high-complexity conditions. The findings support the inclusion of trust-related PSFs as dynamic, experience-sensitive, measurable nodes in probabilistic risk models for automated driving, contributing to more realistic HRA frameworks for ADS safety assessment. This study is part of a broader project aimed at constructing BBN-based HAT models representing how different human profiles (e.g., drivers, operators) and contextual factors (e.g., traffic level, warnings) shape safety outcomes under diverse driving conditions. Future work will combine driver profile characteristics, within-session PSFs, and objective simulation outcomes to develop and validate the complete HAT-BBN.

Acknowledgements

This research was partly supported by University of California Institute of Transportation Studies.

References

[1] SAE On-Road Automated Vehicle Standards Committee, "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles," *SAE Standard J3016_202104*. SAE International, 400 Commonwealth Drive, Warrendale, PA, United States, Apr. 2021. doi: 10.4271/J3016_202104.

[2] R. Gouribhatla and S. S. Pulugurtha, "Drivers' behavior when driving vehicles with or without advanced driver assistance systems: A driver simulator-based study," *Transp. Res. Interdiscip. Perspect.*, vol. 13, p. 100545, Mar. 2022, doi: 10.1016/j.trip.2022.100545.

[3] C. Correa-Jullian, M. Ramos, A. Mosleh, and J. Ma, "Exploring Human-Autonomy Teams in Automated Driving System Operations," in *2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS)*, May 2024, pp. 1–6. doi: 10.1109/ICHMS59971.2024.10555762.

[4] T. A. O'Neill, C. Flathmann, N. J. McNeese, and E. Salas, "Human-autonomy Teaming: Need for a guiding team-based framework?," *Comput. Human Behav.*, p. 107762, Apr. 2023, doi: 10.1016/j.chb.2023.107762.

[5] B. Zhang, J. de Winter, S. Varotto, R. Happee, and M. Martens, "Determinants of take-over time from automated driving: A meta-analysis of 129 studies," *Transp. Res. Part F Traffic Psychol. Behav.*, vol. 64, pp. 285–307, 2019, doi: 10.1016/j.trf.2019.04.020.

[6] J. B. Lyons, K. Sycara, M. Lewis, and A. Capiola, "Human–Autonomy Teaming: Definitions, Debates, and Directions," *Front. Psychol.*, vol. 12, p. 1932, May 2021, doi: 10.3389/fpsyg.2021.589585.

[7] J. Cegarra, H. Unrein, J. Andre, O. Mouton, and J. Navarro, "Driving among autonomous vehicles: The effect of initial trust and driving style on driving behaviors," *Transp. Res. Part F Traffic Psychol. Behav.*, vol. 112, no. March, pp. 99–110, Jul. 2025, doi: 10.1016/J.TRF.2025.03.023.

[8] S. Nordhoff, J. Stapel, X. He, A. Gentner, and R. Happee, "Do driver's characteristics, system performance, perceived safety, and trust influence how drivers use partial automation? A structural equation modelling analysis," *Front. Psychol.*, vol. 14, 2023, doi: 10.3389/FPSYG.2023.1125031.

[9] K. van de Merwe, S. Mallam, and S. Nazir, "Agent Transparency, Situation Awareness, Mental Workload, and Operator Performance: A Systematic Literature Review," *Hum. Factors*, vol. 66, no. 1, pp. 180–208, Jan. 2024, doi: 10.1177/00187208221077804.

[10] C. J. Johnson, M. Demir, N. J. McNeese, J. C. Gorman, A. T. Wolff, and N. J. Cooke, "The Impact of Training on Human–Autonomy Team Communications and Trust Calibration," *Hum. Factors*, Oct. 2021, doi: 10.1177/00187208211047323.

[11] C. Castro *et al.*, "How are distractibility and hazard prediction in driving related? Role of driving experience as moderating factor," *Appl. Ergon.*, vol. 81, p. 102886, Nov. 2019, doi: 10.1016/J.APERGO.2019.102886.

[12] R. Parasuraman and D. H. Manzey, "Complacency and Bias in Human Use of Automation: An Attentional Integration," *https://doi.org/10.1177/0018720810376055*, vol. 52, no. 3, pp. 381–

410, Oct. 2010, doi: 10.1177/0018720810376055.
[13] S. Shahrdar, L. Menezes, and M. Nojoumian, “A survey on trust in autonomous systems,” in *Intelligent Computing*, 2019, vol. 857, pp. 368–386. doi: 10.1007/978-3-030-01177-2_27/TABLES/4.
[14] M. Chu *et al.*, “Work with AI and Work for AI: Autonomous Vehicle Safety Drivers’ Lived Experiences,” *arXiv Prepr. arXiv …*, vol. 1, no. 1, Mar. 2023, doi: 10.1145/3544548.3581564.
[15] F. Vanderhaegen, “Heuristic-based method for conflict discovery of shared control between humans and autonomous systems - A driving automation case study,” *Rob. Auton. Syst.*, vol. 146, p. 103867, Dec. 2021, doi: 10.1016/j.robot.2021.103867.
[16] F. Warg, M. Skoglund, and M. Sassman, “Human Interaction Safety Analysis Method for Agreements with Connected Automated Vehicles,” in *2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*, Sep. 2021, vol. 2021-Septe, pp. 01–07. doi: 10.1109/VTC2021-Fall52928.2021.9625202.
[17] G. Bianchi Piccinini *et al.*, “How Do Drivers Respond to Silent Automation Failures? Driving Simulator Study and Comparison of Computational Driver Braking Models,” *Hum. Factors*, vol. 62, no. 7, pp. 1212–1229, Nov. 2020, doi: 10.1177/0018720819875347.
[18] J. B. Manchon, M. Bueno, and J. Navarro, “How the initial level of trust in automated driving impacts drivers’ behaviour and early trust construction,” *Transp. Res. Part F Traffic Psychol. Behav.*, vol. 86, pp. 281–295, Apr. 2022, doi: 10.1016/j.trf.2022.02.006.
[19] H. Pan, K. Xu, Y. Qin, and Y. Wang, “How does drivers’ trust in vehicle automation affect non-driving-related task engagement, vigilance, and initiative takeover performance after experiencing system failure?,” *Transp. Res. Part F Traffic Psychol. Behav.*, vol. 98, pp. 73–90, Oct. 2023, doi: 10.1016/J.TRF.2023.09.001.
[20] W. Payre *et al.*, “Understanding Drivers’ Trust After Software Malfunctions and Cyber Intrusions of Digital Displays in Automated Cars,” *Hum. Factors Transp.*, vol. 60, pp. 320–328, 2022, doi: 10.54941/AHFE1002463.
[21] C. A. Correa Jullian, “Human-Autonomy Teams in Automated Driving Systems: An Integrated Operational Safety Methodology,” University of California, Los Angeles, 2025. Accessed: Sep. 05, 2025. [Online]. Available: https://escholarship.org/uc/item/6zt6r8r0
[22] R. Xu *et al.*, “The OpenCDA Open-Source Ecosystem for Cooperative Driving Automation Research,” *IEEE Trans. Intell. Veh.*, vol. 8, no. 4, pp. 2698–2711, Apr. 2023, doi: 10.1109/TIV.2023.3244948.
[23] R. Xu, Y. Guo, X. Han, X. Xia, H. Xiang, and J. Ma, “OpenCDA: An Open Cooperative Driving Automation Framework Integrated with Co-Simulation,” *IEEE Conf. Intell. Transp. Syst. Proceedings, ITSC*, vol. 2021-Septe, pp. 1155–1162, 2021, doi: 10.1109/ITSC48978.2021.9564825.
[24] W. Payre, J. Cestac, and P. Delhomme, “Fully Automated Driving: Impact of Trust and Practice on Manual Control Recovery,” *Hum. Factors*, vol. 58, no. 2, pp. 229–241, Mar. 2016, doi: 10.1177/0018720815612319.
[25] W. Payre, J. Cestac, and P. Delhomme, “Intention to use a fully automated car: Attitudes and a priori acceptability,” *Transp. Res. Part F Traffic Psychol. Behav.*, vol. 27, no. PB, pp. 252–263, Nov. 2014, doi: 10.1016/J.TRF.2014.04.009.
[26] A. McKerral, K. Pammer, and C. Gauld, “Supervising the self-driving car: Situation awareness and fatigue during highly automated driving,” *Accid. Anal. Prev.*, vol. 187, p. 107068, Jul. 2023, doi: 10.1016/J.AAP.2023.107068.
[27] B. Metz, J. Wörle, M. Hanig, M. Schmitt, A. Lutz, and A. Neukum, “Repeated usage of a motorway automated driving function: Automation level and behavioural adaption,” *Transp. Res. Part F Traffic Psychol. Behav.*, vol. 81, pp. 82–100, Aug. 2021, doi: 10.1016/j.trf.2021.05.017.
[28] J. Dillmann *et al.*, “Repeated conditionally automated driving on the road: How do drivers leave the loop over time?,” *Accid. Anal. Prev.*, vol. 181, p. 106927, Mar. 2023, doi: 10.1016/j.aap.2022.106927.
[29] X. Han *et al.*, “CDA.AI for OpenCDA: AI pathways for cooperative driving automation research,” *Artif. Intell. Transp.*, vol. 1, p. 100002, Jul. 2025, doi: 10.1016/j.ait.2025.100002.