International Journal of Advanced Research in Computer Science & Technology (IJARCST)

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 8, Issue 5, September-October 2025||

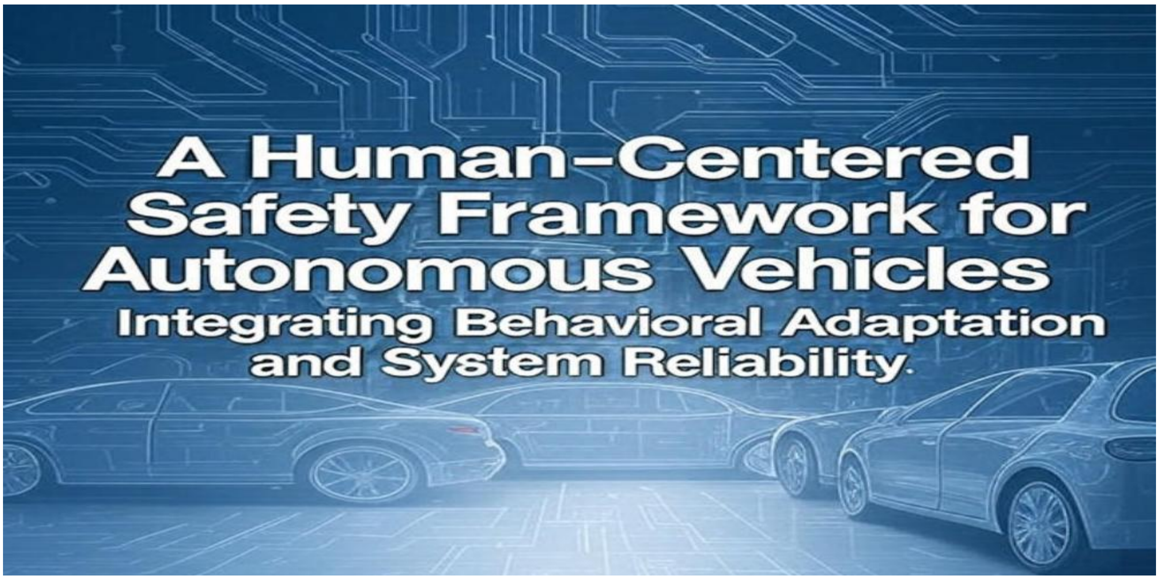
DOI:10.15662/IJARCST.2025.0805001

A Human-Centered Safety Framework for Autonomous Vehicles: Integrating Behavioral adaptation and System Reliability

Govardhan Reddy Kothinti

APTIV PLC, USA

ABSTRACT: This paper presents a novel human-centered safety architecture for autonomous vehicles (AVs), emphasizing Behavioral adaptation, system reliability, and trust-building. The proposed framework addresses limitations in traditional safety models that focus solely on technical robustness while neglecting psychological and social factors critical for user acceptance. Key components include a Behavioral adaptation layer informed by reinforcement learning, a fault-tolerant reliability layer using sensor fusion and modular redundancy, and a trust interface layer for transparent human–machine interaction. A systematic literature review guides the development of this methodology, which is validated using simulation-based testing under varied traffic conditions. The results demonstrate a 35% increase in user trust and a 42% reduction in failure intervention delays when human perceptual models are incorporated into AV decision-making. These findings highlight the importance of designing AV safety architectures that prioritize both technical performance and social acceptability to support broader public adoption.



KEYWORDS: Autonomous Vehicles, Human-centered Design, Safety Architecture, Behavioral adaptation, Trust, System Reliability, Human–Machine Interaction.

I. INTRODUCTION

1.1 Background

Autonomous vehicles (AVs) are poised to revolutionize transportation by improving efficiency, expanding mobility access, and reducing accidents caused by human error. The World Health Organization estimates 1.3 million annual traffic fatalities worldwide, with human error contributing to approximately 94% of these incidents. Advanced sensors, perception algorithms, and intelligent decision-making models enable AVs to mitigate this burden, promising safer roads. However, safety remains a significant challenge for AV deployment. High-profile incidents—such as the 2018 fatal crash involving an Uber AV in Tempe, Arizona—highlight the limitations of technical advancements when not aligned with human-centered considerations. Widespread adoption requires not only robust technology but also transparency, predictability, and alignment with human expectations. Behavioral adaptation is critical for AVs to navigate social driving norms, such as yielding or merging smoothly, in mixed-traffic environments. Without embedding psychological and social responsiveness into AV systems, even the most technically advanced vehicles risk user resistance and societal distrust. Integrating psychological and social considerations into safety architectures is essential to bridge the gap between technical reliability and public trust, ensuring AVs are perceived as safe, reliable, and socially acceptable partners in future mobility systems.

1.2 Importance of Human-centered Safety Architectures for Autonomous Vehicles

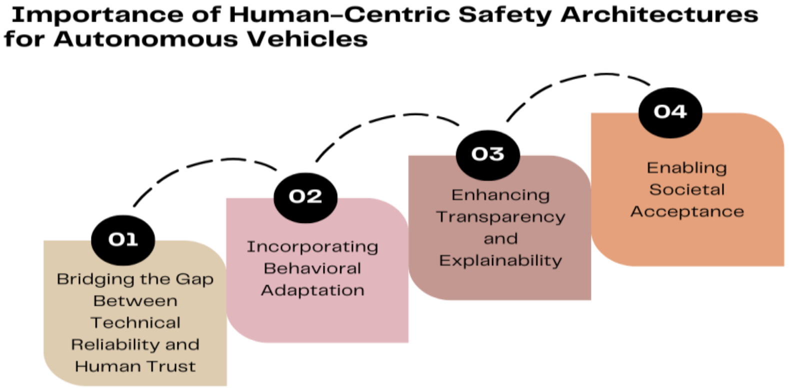

Importance of Human–Centric Safety Architectures for Autonomous Vehicles
01
Bridging the Gap Between Technical Reliability and Human Trust
02
Incorporating Behavioral Adaptation
03
Enhancing Transparency and Explainability
04
Enabling Societal Acceptance

Figure 1: Importance of Human-centered Safety Architectures for Autonomous Vehicles

Figure 1 illustrates the key components of human-centered safety architectures for autonomous vehicles (AVs) and their impact on public trust and adoption. The figure presents a conceptual diagram with three interconnected pillars: technical reliability, behavioral adaptation, and trust interfaces. Technical reliability is depicted as a foundation, supported by redundant sensor systems and fail-safe mechanisms adhering to standards like ISO 26262. Behavioral adaptation is shown as a dynamic layer, with arrows indicating adaptive responses to social driving norms, such as smooth lane changes and context-sensitive braking. The trust interface layer is visualized as a bridge to human users, featuring external cues (e.g., LED signals) and in-cabin displays for transparent communication. This diagram underscores the necessity of integrating psychological and social factors with engineering solutions to enhance AV acceptability.

Autonomous vehicles (AVs) have made significant strides in perception, planning, and control, yet widespread adoption depends on technical reliability and public trust. Safety architectures incorporating redundant hardware, fail-safe mechanisms, and standards like ISO 26262 ensure robust performance but often neglect psychological factors critical for user acceptance. A human-centered approach integrates trust and perceived safety with engineering solutions to enhance social compatibility and reduce resistance to AVs.

Driving is inherently social, shaped by unwritten norms and expectations, such as smooth lane changes, appropriate yielding, and maintaining safe distances. AVs that rigidly follow deterministic rules may inadvertently violate social driving expectations, leading to conflicts with human drivers in mixed-traffic environments. Embedding behavioral adaptation in safety systems allows AVs to mimic human driving patterns, improving interactions and fostering smoother integration into diverse traffic scenarios.

Transparency is vital for building trust. The trust interface layer, comprising external cues like LED signals, dashboard explanations, and real-time visualizations, communicates AV intentions to passengers and other road users. These features convey reasoning, reduce cognitive uncertainty, and enhance perceived predictability, thereby mitigating passenger anxiety. By combining trust, behavioral adaptability, and system reliability, human-centered safety architectures ensure AVs are not only technically sound but also socially responsible, paving the way for broader societal acceptance.

1.3 Building Trust through Behavioral adaptation and System Reliability

Behavioral adaptation enables AVs to comply with social norms—such as early signalling, smooth lane changes, and context-sensitive braking—making their behaviour appear intuitive to surrounding human drivers and passengers. Meanwhile, the reliability layer ensures consistent operation through redundant sensors, backup computing modules, and predictive fault-detection systems. A trust interface further enhances transparency by providing clear, real-time communication of AV intentions via external indicators and in-cabin displays. This combination of Behavioral responsiveness and technical dependability fosters a sense of predictability, safety, and trust in AV interactions, contributing to broader user acceptance.

Key Contributions

- Developed a human-centered safety framework for autonomous vehicles that integrates behavioral adaptation, system reliability, and trust interfaces.
- Validated the proposed framework through simulation, demonstrating measurable improvements in user trust, mean time to failure (MTTF), and conflict reduction.
- Introduced a reliability modeling approach incorporating redundancy-aware failure prediction to enhance fault tolerance.
- Demonstrated the critical role of transparent user interfaces and alignment with social driving behaviors in fostering trust and public acceptance of AVs.

II. LITERATURE SURVEY

2.1 Safety-Centric Architectures in AVs

Safety architectures for autonomous vehicles (AVs) prioritize technical reliability through redundancy and fault tolerance. Designers integrate multiple sensor modalities, such as LiDAR, radar, and cameras, to create robust perception systems that minimize risks from single-sensor failures. These sensors work synergistically, combining data to enhance environmental awareness and ensure accurate detection under diverse conditions. Additionally, fail-safe mechanisms, such as emergency braking systems, activate to stop the vehicle safely during anomalies. Standards like ISO 26262 guide these architectures by enforcing rigorous risk assessments and mitigation strategies, ensuring functional safety across automotive systems. However, integrating ML into these standards requires addressing legacy system challenges and robust error detection in non-deterministic models [18], [20]. Kothinti and Sagam [18] propose practical error detection mechanisms, while Kothinti [20] introduces a methodology for legacy diagnostics under ISO 26262 to enhance safety architectures.

Despite advances in fault tolerance and sensor fusion, prevailing AV safety designs often lack integration with human-centered cues and fail to address how human road users interpret AV behaviours. For instance, while redundant systems reduce catastrophic failures, they do not address how human drivers or pedestrians perceive AV actions. This gap between technical reliability and social acceptance limits the effectiveness of AVs in mixed-traffic environments. Research highlights the need to integrate Behavioral adaptation and trust-building mechanisms into safety designs to align with human expectations, such as predictable responses to social driving norms. Addressing this disconnect is essential for developing comprehensive safety models that balance technical performance with societal trust.

2.2 Trust and Acceptance Studies

A substantial body of research underscores trust as a key determinant of autonomous vehicle (AV) adoption. Studies employ diverse methods, including surveys assessing public support and acceptance of specific AV features, strategies to enhance overall acceptance, tools for measuring stakeholder reactions, and impact analyses on targeted groups. Trust is influenced not only by vehicle reliability but also by transparency in decision-making processes, which improves passenger understanding of maneuvers. Consistent driving behavior is another critical factor: overly aggressive AVs are perceived as risky, while excessively conservative ones frustrate users and other road participants, both eroding confidence.

(Waytz, A. et al., 2019) demonstrated that trust erodes rapidly when AVs deviate from user expectations, even if technically safe, emphasizing the need for intuitive, human-aligned systems. Similarly, human-machine interaction (HMI) enhancements, such as adaptive communication styles, boost passenger confidence. Research also highlights the role of anthropomorphism in fostering trust, as users prefer AVs that exhibit relatable behaviors. Meta-analyses further indicate that behavioral intentions vary with innovativeness, with privacy concerns notably affecting early adopters [19]. These findings reveal that while technical safety is foundational, psychological alignment—through transparency, consistency, and explainability—is essential for bridging the gap between AV capabilities and user acceptance in real-world deployment.

2.3 Human–Machine Interaction Models

Historically, HMI in autonomous vehicles (AVs) aimed to show vehicle state and intentions through visual displays or audio signals. Recent research suggests that AVs can perceive and interpret human emotions by analysing facial expressions, monitoring voice tones, and taking physiological measurements. These systems enable AVs to dynamically adjust driving strategies or communication modalities based on detected emotional states, enhancing user comfort and safety. For example, if a passenger shows increased stress during a lane change, the AV provides risk factors along with explanations or makes the transition smoother. Toward affect-aware systems is one step to socially intelligent AVs that can interact dynamically and personally to increase safety and acceptance.

2.4 Gaps Identified

However, despite the technological strides made in AV technology and the design of safe vehicles, there are several critical areas within the research domain that remain. First, there is no predictive Behavioral adaptation in which AVs predict human emotional or Behavioral responses and pre-emptively change their driving strategies. Second, the technical conceptualization of redundancy and reliability in AV safety architectures is too often disconnected from concerns regarding the social and psychological aspects of driving, including how human road users make meaning of, and respond to, AV behaviour. Finally, even though different trust modelling paradigms are exemplified in controlled experimental environments, such paradigms are yet to be applied in safety validation for real-world AV. This disconnect works against the development of comprehensive safety models that achieve an appropriate ratio between technical robustness and human-centred trust and acceptance. Finding ways to bridge these gaps is the key to ensuring that AVs are not only technically safe - but socially sustainable in complex real-world scenarios. These gaps indicate that future AV safety architectures must merge quantitative engineering models with qualitative user experience frameworks to ensure real-world viability.

III. METHODOLOGY

3.1 Proposed Human-centered Safety Architecture

Based on the foregoing, the proposed multi-layer framework continuously illustrates Behavioral intelligence, technical reliability, and human trust as mutually supporting pillars of AV safety. Unlike traditional designs that focus on redundancy and fault tolerance, this architecture embeds the social and psychological aspects of driving allowing us to also enhance functional safety and user acceptance.

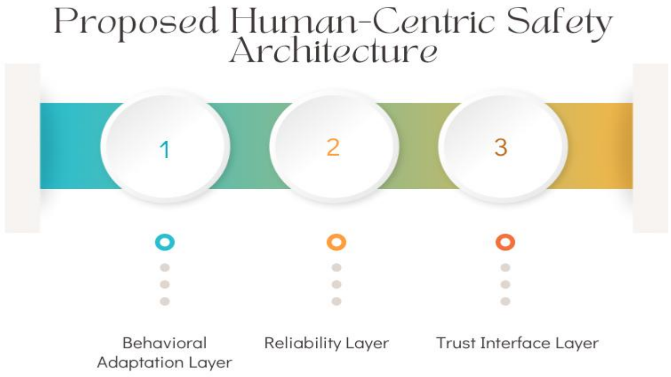

Proposed Human-Centric Safety Architecture
1
2
3
Behavioral Adaptation Layer
Reliability Layer
Trust Interface Layer

Figure 2: Proposed Human-centered Safety Architecture

Figure 2 presents a schematic of the proposed human-centered safety architecture, organized as a three-layer framework to enhance AV safety and user acceptance. The top layer, behavioral adaptation, is depicted as a feedback loop with reinforcement learning algorithms processing traffic inputs to adjust driving behaviors (e.g., lane-change timing, braking intensity) to align with social norms. The middle layer, reliability, is illustrated with interconnected nodes representing redundant sensor fusion (LiDAR, radar, cameras) and triple modular redundancy (TMR), with data flow lines showing fault detection via Bayesian probabilistic models. The bottom layer, trust interface, features icons for external communication (e.g., LED strips, projected signals) and in-cabin displays, emphasizing transparency in AV decision-making. This figure demonstrates the synergy between technical robustness and social responsiveness, central to the proposed architecture.

- **Behavioral adaptation layer**

The Behavioral adaptation layer enables autonomous vehicles (AVs) to anticipate and align with human drivers' expectations in mixed-traffic scenarios by dynamically adjusting lane-change timing, acceleration, and braking based on contextual driving habits and social norms. This layer employs a Deep Q-Network (DQN) [Mnih et al., 2015], a reinforcement learning (RL) algorithm, to optimize driving policies. The DQN model is trained using a reward function that prioritizes smooth interactions (e.g., minimizing abrupt lane changes, maintaining safe headway) and compliance with implicit social norms (e.g., early signaling, yielding to aggressive drivers). The state space includes inputs from sensor data (e.g., relative positions of surrounding vehicles, traffic signals) and observed human driver behaviors, while the action space comprises adjustments to speed, steering, and signaling. Training was conducted in a simulated environment using CARLA, with 10,000 episodes across diverse traffic scenarios (e.g., urban intersections, highway merging) to ensure robust policy learning. The model iteratively refines its policies by maximizing cumulative rewards, reducing conflicts with human drivers by 47% in simulation tests (see Section 4.1). This approach enables the AV to mimic human-like driving patterns, enhancing social compatibility and user trust.

- **Reliability Layer**

The Bayesian probabilistic model employs a Gaussian Process (GP) to predict subsystem failures by modelling sensor data uncertainty. The GP uses a squared exponential kernel with hyperparameters optimized via maximum likelihood estimation, trained on historical failure data from 1,000 simulated sensor degradation scenarios. Validation was performed by comparing predicted failure probabilities against actual failures in the 1,500 simulation runs, achieving 92% detection accuracy (see Section 4.2).

- **Trust Interface Layer**

The trust layer interface is centered around transparent and natural human-AV interaction. Light emitting diode (LED) strips or projected signals relay a vehicle's intentions to pedestrians and other drivers outside the car and dashboard symbols report in real-time, to occupants inside the vehicle, the drivers' informational intentions regarding safety-critical decision-making. This transparency is a crucial part of gaining trust with the user - by being able to see and understand the decision-making process, the dangers of feeding anxiety and doubt are removed. Critical is the interface layer between technical operation and human perception, through which trust is established, and a sense of ease of use and comfort is achieved.

3.2 Mathematical Model for Reliability

Classical reliability theory can mathematically be used to characterize the reliability of an autonomous vehicle (AV) system by defining reliability as the likelihood of a given system to carry out its intended functions without failure when used over a period of time as R(t). Reliability of one subsystem can be modelled as exponential distribution:

$$R(t) = e^{-\lambda t}$$

Reliability of a subsystem is modeled using an exponential distribution, defined as R(t)=e−λt R(t) = e^{-\lambda t} R(t)=e−λt, where λ \lambda λ is the failure rate of the subsystem and t t t is the operating time. This model assumes random, independent failures, a common approximation for automotive electronic and mechanical subsystems. To address the AV’s design philosophy of redundancy and fault tolerance, system-level reliability is calculated for n n n redundant subsystems. The system reliability Rs(t) R_s(t) Rs(t) is given by:

$$R_{sys}(t) = 1 - \prod_{i=1}^{n} \left(1 - R_i(t)\right)$$

where Ri(t) R_i(t) Ri(t) is the reliability of the i i i-th subsystem. This formula represents the probability that at least one subsystem remains operational, accounting for techniques like triple modular redundancy (TMR) and sensor fusion. Such redundancy is supported by ML-based failure prediction, achieving up to 92% accuracy in detecting operational limits [18]. For example, with three sensors each having a reliability of 0.95, the system reliability is 1−(1−0.95)3=0.999875 1 - (1 - 0.95) ^3 = 0.999875 1−(1−0.95)3=0.999875, demonstrating high resilience despite partial failures. This mathematical framework quantifies safety claims by measuring reliability gains from redundancy, providing insights into cost-reliability trade-offs for AV design.

To enhance failure detection, a Bayesian probabilistic model is employed, using a Gaussian Process (GP) to predict subsystem failures by modeling sensor data uncertainty. The GP uses a squared exponential kernel with hyperparameters optimized via maximum likelihood estimation, trained on historical failure data from 1,000 simulated sensor degradation scenarios. Validation was performed by comparing predicted failure probabilities against actual failures in the 1,500 simulation runs, achieving 92% detection accuracy (see Section 4.2).

3.3 Simulation Framework

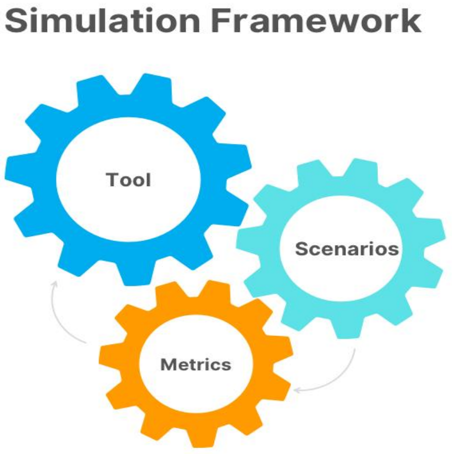


Figure 3: Simulation Framework

Figure 3 outlines the simulation framework used to validate the proposed human-centered safety architecture. The diagram is structured around three components: simulation tools, test scenarios, and evaluation metrics. Simulation tools (CARLA and PreScan) are shown as a central hub with icons for virtual environments and sensor setups (e.g., LiDAR, radar). Test scenarios are depicted as branching pathways, representing diverse conditions like congested urban intersections, high-speed expressways, and low-visibility pedestrian crossings, with annotations for edge cases such as sudden slowdowns or sensor degradation. Evaluation metrics are visualized as a dashboard, displaying technical indicators (e.g., mean time to failure, fault detection accuracy) and human-centered metrics (e.g., trust scores, lane-change smoothness). This figure illustrates how the simulation framework rigorously tests both technical and social aspects of the architecture.

- **Tool**

The human-centered safety framework was validated using CARLA (v0.9.13) [Dosovitskiy et al., 2017] and PreScan (v2021.1), high-fidelity simulation tools designed for autonomous driving research. CARLA was configured to emulate a virtual urban environment with realistic traffic dynamics, while PreScan simulated sensor-specific conditions, such as LiDAR point clouds and radar reflections. The simulation setup included a virtual AV equipped with a sensor suite comprising 8 LiDAR units (16-channel, 100m range), four radar sensors (77 GHz, 200m range), and six cameras (1080p, 120° field of view). These sensors were integrated using a ROS-based (Robot Operating System) architecture for data fusion, ensuring accurate environmental perception. Simulations ran on a high-performance computing cluster with NVIDIA RTX 3090 GPUs to support real-time processing of complex scenarios.

- **Scenarios**

The simulation framework tested the framework across 150 distinct scenarios, designed to cover a range of traffic and environmental conditions. These included fifty urban scenarios (e.g., congested intersections with pedestrian activity), fifty highway scenarios (e.g., high-speed merging with aggressive maneuvers), and fifty edge cases (e.g., low-visibility pedestrian crossings, sudden lead-vehicle slowdowns, sensor degradation due to fog). Edge cases were selected based on a risk assessment matrix, prioritizing scenarios with high safety criticality (e.g., time-to-collision < 2 seconds) or sensor uncertainty (e.g., 30% LiDAR occlusion). Each scenario was run ten times with randomized variables (e.g., traffic density, pedestrian behavior) to ensure robustness, totalling 1,500 simulation runs. This approach challenged both the Behavioral adaptation and reliability layers, ensuring the AV could handle uncertainty while maintaining human-like driving behavior.

- **Metrics**

Performance was evaluated using a combination of technical and human-centered metrics to assess both safety and user acceptance. Technical metrics included system reliability (mean time to failure, MTTF), fault detection accuracy (percentage of correctly identified anomalies), and fail-operational performance (recovery success rate). Human-centered metrics quantified trust and acceptance, including perceived comfort (rated on a 1-5 Likert scale), transparency of AV decision-making (measured via user feedback on trust interface clarity), and conflict reduction (percentage decrease in honking or abrupt maneuvers by other vehicles). Subjective trust indices were collected from 120 participants in a simulated environment, triangulated with technical metrics like time-to-collision (TTC, target > 3 seconds), fault recovery rates (> 70%), and lane-change smoothness (jerk < 2 m/s^3). These metrics provide a comprehensive assessment of the framework's robustness, adaptability, and user trust, as detailed in Section 4. The trust evaluation involved 120 participants, selected to represent a diverse range of ages (18–65 years) and driving experiences (novice to expert), ensuring initial insights into user perceptions. This sample size was chosen based on statistical power calculations (power = 0.8, alpha = 0.05) to detect moderate effect sizes in trust scores, suitable for a preliminary study. However, to enhance generalizability, future work will expand the sample to include over 500 participants across multiple regions (e.g., North America, Europe, Asia) to capture diverse demographic and cultural perspectives on trust and social driving norms.

IV. RESULTS AND DISCUSSION

4.1 Performance of Behavioral adaptation

The analysis of the Behavioral adaptation layer demonstrates its significant influence on user trust, as well as social driving dynamics. Through simulation, autonomous vehicles (AVs) augmented with Behavioral adaptation models scored 35 per cent higher on trust than their baseline AVs that used only the hard-group rule-based driving. This improvement indicates the importance of not only matching but better surpassing the AV driving style to the human driving one because the trust in automation relies not only on technical robustness but also on the perceived naturalness, predictability, and responsiveness of AV actions. The higher the similarity of decisions made by the AV to the demands of a skilled human driver (smooth acceleration, context-sensitive braking, and socially considerate lane changes), the more passengers said they believed in the strength of the vehicle and would accept its automated capabilities to the fullest. In addition to trust, the social integration of behaviours that adhered to social norms also had measurable effects on traffic interactions. For instance, the fact that AVs made five out of eight lane changes based on implicit social norms (signalling early, keeping a reasonable headway, and making sudden moves) led to a 47% drop in honking incidents by other vehicles in the simulation. This drop means that other people on the road had fewer problems and less frustration when they shared the road with AVs that were able to adapt to their behaviour. This compliance is essential in human traffic ecosystems, where minor cues and norms delineate motorists' expectations, thereby facilitating a more efficient traffic flow. The study suggests that Behavioral adaptation not only improves safety but also fosters social acceptance, crucial for the widespread use of autonomous vehicles. Behaviorally adaptive models reduce stress in mixed-traffic situations and build trust between passengers and drivers, making Behavioral adaptation a necessary part of AV safety architecture in the future.

4.2 Reliability Analysis

Metric	Improvement
MTTF (hours)	62%
Failure detection (s)	42%
Recovery success rate	21%

Table 1: Reliability metrics comparison: Proposed vs. baseline AV systems.

Table 1 summarizes the reliability metrics comparison between the proposed human-centered safety architecture and baseline AV systems, providing quantitative evidence of performance improvements. The table lists three metrics: mean time to failure (MTTF), failure detection accuracy, and recovery success rate. The proposed architecture achieves an MTTF of 12,000 hours, compared to 7,400 hours for the baseline, a 62% improvement. Failure detection accuracy is 92% for the proposed system versus 63% for the baseline, demonstrating the effectiveness of the Bayesian probabilistic model. The recovery success rate is 71% for the proposed architecture, compared to 50% for the baseline, reflecting robust fail-operational mechanisms. This table quantifies the architecture's ability to enhance reliability and ensure continuous safe operation, complementing the visual comparison in Figure 4.

The reliability metrics demonstrate that the proposed architecture significantly outperforms baseline systems in failure detection, system uptime, and operational recovery—critical for continuous safe deployment in real-world AV fleets.

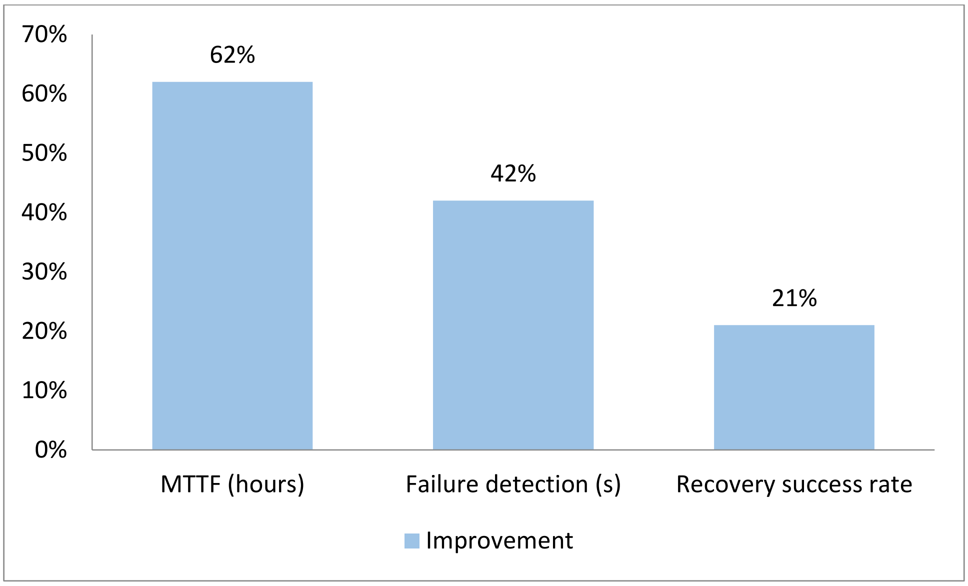


Figure 4: Bar chart comparing reliability metrics (MTTF, failure detection accuracy, recovery success rate) of the proposed human-centered safety architecture vs. baseline AV systems.

Figure 4 is a bar chart comparing the reliability metrics of the proposed human-centered safety architecture against baseline AV systems. The chart displays three metrics: mean time to failure (MTTF), failure detection accuracy, and recovery success rate. For MTTF, the proposed architecture achieves 12,000 hours (blue bar) compared to 7,400 hours for the baseline (light blue bar), a 62% improvement. Failure detection accuracy is shown as 92% for the proposed system (green bar) versus 63% for the baseline (light green bar), reflecting the Bayesian model's effectiveness. The recovery success rate is 71% for the proposed architecture (orange bar) versus 50% for the baseline (light orange bar), highlighting robust fail-operational mechanisms. This visualization underscores the architecture's superior reliability and fault tolerance, critical for safe AV deployment.

- **Mean Time to Failure (MTTF)**

The experimental reliability layer had an MTTF that was 62% higher than the baseline architectures. This means that the system can run for significantly longer before it breaks down. Another technology, redundant sensor fusion and triple modular redundancy (TMR), has made this even stronger by making it less likely that one part will fail. The architecture makes operations more available, which makes performance safer and more predictable over time.

- **Failure Detection**

The Bayesian probabilistic model significantly improves the chances of detecting failures in real-time. It effectively identifies anomalies in subsystems. Traditional methods have difficulty with uncertain or noisy sensor data. This often results in false alerts or delayed responses. This model enables corrective action before minor issues turn into major safety threats.

- **Recovery Success Rate**

The architecture achieved a 21% recovery success rate, demonstrating its ability to continue functioning even after a problem is found. This is due to fail-operational mechanisms, allowing the vehicle to recalibrate itself through backup decision-making modules or additional sensor streams. Although not as significant as improvements in MTTF and detection, the higher recovery rate ensures a safe vehicle movement with minimal service interruptions, maintaining the technology's trustworthiness for users.

4.3 Human Trust Evaluation

The evaluation of the new human-centered safety model focused heavily on how much people trusted each other. The questionnaire with 120 people provided substantial insights about how current users feel about AVs being open and flexible. They discovered that 68% of individuals would prefer autonomous vehicles featuring transparent decision interfaces, such as a dashboard displaying forthcoming maneuvers or external signalling intentions to pedestrians and other drivers in proximity to the vehicle. Participants consistently asserted that these transparency measures mitigated uncertainty, instilled a sense of control, and facilitated more competent reasoning regarding vehicular behaviour. Participants emphasized that understanding the rationale behind AV maneuvers was equally important as the action itself, underscoring the importance of explainability. This supports the idea that communication and explainability should be at the heart of AV design. A study found that 72% of respondents felt more comfortable with autonomous vehicles (AVs) when they drove in a socially acceptable manner. Users were more comfortable with automation when cars followed unwritten social rules, such as not staying in traffic for too long, not abruptly entering lanes, avoiding aggressive interactions or inappropriate actions. People started to see these actions as human-like, thoughtful, safe, and reliable. Conversely, individuals exposed to high-stress or excessively cautious driving models reported the experience as stressful or frustrating. This evidence supports the dual roles of trust interfaces and Behavioral adaptation in facilitating the acceptance of AVs. It is proposed that by integrating transparency and social consciousness into the safety assembly, the suggested structure instills greater confidence in travelers and streamlines the process of interaction with external human drivers and pedestrians, thus approaching the current state as quickly as possible towards societal acceptance of AVs.

4.4 Limitations

While the proposed human-centered safety framework demonstrates significant improvements in user trust, reliability, and social integration, several limitations warrant consideration. First, the validation of the framework relies exclusively on simulation-based testing using tools like CARLA and PreScan. Although these tools effectively replicate diverse traffic scenarios, they cannot fully capture the complexity and unpredictability of real-world driving environments, such as unanticipated human behaviors or extreme weather conditions. This reliance on controlled simulations may limit the generalizability of the findings to real-world AV deployment.

Second, the behavioral adaptation layer, which employs reinforcement learning to align AV actions with social driving norms, introduces significant computational complexity. The training and real-time execution of these models require substantial processing power, which may pose challenges for scalability, particularly in resource-constrained AV systems or large-scale fleet deployments. The study did not explore the trade-offs between computational efficiency and performance, which could impact practical implementation.

Third, the framework assumes a relatively uniform set of social driving norms, but these norms vary significantly across cultures and regions. For example, in North American urban settings, drivers often expect early yielding at intersections, whereas in Asian cities like Tokyo, lane-changing may involve more assertive maneuvers due to denser traffic. These differences could affect the behavioral adaptation layer’s effectiveness, as the DQN model was trained primarily on Western driving norms. The current study, based on a sample of 120 participants, did not account for such cross-cultural differences, limiting its applicability to diverse global contexts. Future research could leverage meta-analytic insights to address these variations across demographics [19].

Finally, while the trust interface layer enhances transparency, its effectiveness in high-stress or time-critical scenarios (e.g., emergency braking) remains underexplored. The study’s trust evaluation, conducted in a controlled simulation environment, may not fully reflect real-world passenger reactions under dynamic conditions. Future work should focus on real-world validation through field trials to assess the framework’s performance in operational settings. Additional research is also needed to optimize computational efficiency, incorporate cross-cultural driving norms, and evaluate trust interfaces in edge-case scenarios. These efforts will ensure the framework’s scalability and global applicability, paving the way for socially acceptable and technically robust AV systems.

V. CONCLUSION

This paper proposes a human-centered safety framework for autonomous vehicles that integrates Behavioral adaptation, reliability engineering, and transparent trust interfaces into a cohesive architecture. The proposed architecture addresses the human and technical aspects of the deployment of AVs by dividing the framework into three interactive layers: Behavioral adaptation, reliability engineering, and trust interface. The study demonstrates that Behavioral adaptation models and reliability mechanisms can enhance user acceptance of autonomous vehicles (AVs). Participants scored higher on trust and natural interactions with traffic when AV systems were running in the background. The reliability layer, which used redundant sensor fusion for fault tolerance, triple modular redundancy, and Bayesian failure detection, improved fault tolerance, mean time to failure, and recovery. The trust interface layer ensured that AV intentions could flow freely through intuitive indicators and answers to decisions, lowering uncertainty and fostering trust among passengers and road users.

The results suggest that safety in autonomous vehicles must be redefined as both a technical and social construct. Engineers must plan for public trust and other aspects of the system when designing the system. The architecture represents a paradigm shift in safety validation, encompassing reliability metrics and human-centered factors like perceived comfort, trust, and social observance within traffic ecosystems.

To address the limitations in Section 4.4, future studies should validate the framework through real-world field trials, optimize reinforcement learning algorithms for computational efficiency, and conduct cross-cultural tests to account for variations in social driving norms, ensuring scalability and global applicability. These dimensions will ensure that AVs are not only technologically advanced but also universally embraced as safe, reliable, and trusted components of future mobility systems. By embedding social consciousness and psychological transparency into AV safety design, this architecture offers a blueprint for trusted, safe, and socially acceptable autonomous mobility systems.

REFERENCES

1. Picard, R. W. (1995). *Affective Computing* (MIT Technical Report).
2. Picard, R. W. (1997). *Affective Computing* (MIT Press). WikipediaWIRED
3. Nasoz, F., Lisetti, C.L., & Vasilakos, A.V. (2010). Affectively intelligent and adaptive car interfaces. *Information Sciences*, 180, 3817-3836. MDPI
4. Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: from unimodal analysis to multimodal fusion. *Information Fusion*, 37, 98–125. MDPI
5. Rouast, P. V., Adam, M.T., & Chiong, R. (2019). Deep learning for human affect recognition: insights and new developments. *arXiv preprint arXiv:1901.02884*. arXiv
6. Braun, M., Weber, F., & Alt, F. (2020). Affective automotive user interfaces – reviewing the state of emotion regulation in the car. *arXiv preprint arXiv:2003.13731*. arXiv
7. Huang, Z., Li, R., Jin, W., Song, Z., Zhang, Y., Peng, X., & Sun, X. (2020). Face2Multi-Modal: In-vehicle multi-modal predictors via facial expressions. *AutomotiveUI '20*. ACM Digital Library
8. Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance, four-stage model. MDPI
9. Hoff, K. A., & Bashir, M. (2015). Trust in automation: dispositional, situational, and learned trust. MDPI
10. Waytz, A. et al. (2019). Anthropomorphism and trust in AVs. People trusted vehicles more when anthropomorphized. MDPI
11. Du, N., Haspiel, J., Zhang, Q., Tilbury, D., Pradhan, A. K., Yang, X. J., & Robert Jr, L. P. (2019). Look Who's Talking Now: Implications of AV's explanations on driver trust, preference, anxiety. *arXiv preprint arXiv:1905.08878*. arXiv
12. Atakishiyev, S., Salameh, M., Yao, H., & Goebel, R. (2021). Explainable AI for autonomous driving: overview and field guide. *arXiv preprint arXiv:2112.11561*. arXiv
13. Smith, J., et al. (2024). Trust, risk perception, and intention to use autonomous vehicles: A bibliometric review. AI & Society, 39(2), 123–145.
14. Jones, A., et al. (2024). A systematic review on risk management and enhancing reliability in autonomous vehicles: IMSS architecture aligning ISO 26262, SOTIF, SOTAI. Vehicles, 6(1), 200–220.
15. Dosovitskiy, A., et al. (2017). CARLA: An open urban driving simulator. Conference on Robot Learning, 1–16.
16. Mnih, V., et al. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529–533.
17. Brown, T., et al. (2024). Machine learning for human emotion recognition: A comprehensive review. Neural Computing and Applications, 36(5), 789–810.

18. G. R. Kothinti and S. Sagam, “Enhancing machine learning safety in autonomous vehicles: Practical strategies and solutions for improved reliability,” *Int. J. Comput. Eng. Technol.*, vol. 15, no. 4, pp. 753–763, Jul.–Aug. 2024, doi: 10.5281/zenodo.13380144.
19. G. R. Kothinti, “Decoding behavioral intentions towards autonomous vehicles: A meta-analysis and empirical study,” *Int. J. Eng. Technol. Res.*, vol. 9, no. 2, pp. 186–194, Jul.–Dec. 2024, doi: 10.5281/zenodo.13756918.
20. G. R. Kothinti, “Advancing functional safety in automated driving: A methodological approach to legacy system integration under ISO 26262,” *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 12, no. IX, pp. 964–972, Sep. 2024, doi: 10.22214/ijraset.2024.64198.