

Perspective

Trustworthy Companion AI for Human-Aware Transition of Control: Motivation, Architecture, and Research Roadmap

Roberta Presta ^{1,*}, Flavia De Simone ¹, Lorenzo Bacchiani ² and Roberto Girau ²

¹ Centro Scienza Nuova, Università degli Studi Suor Orsola Benincasa, 80135 Naples, Italy; flavia.desimone@unisob.na.it

² Department of Computer Science and Engineering, University of Bologna, 40136 Bologna, Italy; lorenzo.bacchiani2@unibo.it (L.B.); roberto.girau@unibo.it (R.G.)

* Correspondence: roberta.presta@unisob.na.it

Abstract

Transitions of control between automated driving systems and human drivers remain safety-relevant and cognitively demanding moments in human–automation interaction. Recent studies show that transition performance depends not only on takeover timing or response speed but also on traffic complexity, driver readiness, automation limitations, trust calibration, and situational-awareness recovery. As in-vehicle interaction evolves toward conversational and agentic AI assistance, takeover support also becomes a problem of governing how natural-language AI systems communicate with the driver under uncertainty. This paper proposes a digital-twin-mediated framework for human-aware takeover support in automated driving. In this framework, the companion AI is treated as an assumed LLM-based in-vehicle conversational or agentic assistant used as an advisory interaction component. The contribution is defined at the architectural level: human, vehicle, and context/road digital twins provide structured semantic state abstractions through a semantic state interface exposing confidence, freshness, provenance, and consistency metadata, while a trustworthy companion AI (TCAI) layer grounds, constrains, validates, and governs companion AI output proposals before HMI delivery. Building on the research on driver-state monitoring, adaptive HMI, trust calibration, explainability, conversational assistance, and human assistance systems (HASs), the framework coordinates advisory interaction across vigilance support, contextual explanation, trust-calibrating communication, and directive handover guidance. The TCAI layer combines bounded reasoning, human-factor-derived guardrails, state-consistency management, dynamic explanation-depth control, trust-dynamics modeling, graded watchdog veto handling, mandatory access-control assumptions, and deterministic fallback. Safety-critical vehicle-control and minimum risk condition (MRC) functions remain assigned to the deterministic vehicle-control stack, while the authorized output path of the TCAI layer is validated HMI delivery. The paper concludes with a validation agenda and technical roadmap covering planned transitions, urgent handovers, degraded or adversarial conditions, temporal fusion of driver-state evidence, phase-sensitive HMI policies, trust-calibration trajectories, driver veto and partial-disabling mechanisms, and staged simulator-to-vehicle evaluation. Although motivated by SAE Level 3 automation, the framework may also inform fallback-related Level 4 scenarios in which human and automated agency must be managed under uncertainty.



Academic Editor: Valeri Mladenov

Received: 21 May 2026

Revised: 19 June 2026

Accepted: 22 June 2026

Published: 24 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: digital twin; human–AI interaction; takeover support; automated driving; trustworthy companion AI; human assistance systems; driver state estimation; hybrid AI architecture; runtime governance; trust calibration

1. Introduction

Conditional automation at SAE Level 3 [1] places human drivers in a structurally ambiguous role: formally relieved of the driving task during automated operation yet remaining the designated fallback when the system reaches the boundary of its operational design domain (ODD) or encounters a fault. At that boundary, a takeover request (TOR) is issued and the driver is expected to regain situational awareness (SA) and resume vehicle control, often in only a few seconds [2].

The fragility of this arrangement has been extensively documented, particularly as automation transforms the driver from an active operator into a passive supervisor [3,4], leading to an out-of-the-loop performance problem in which the human agent struggles to rapidly reconstruct system status and environmental context [5,6]. Reaction times to TORs are highly variable and often increase substantially when drivers are engaged in non-driving-related tasks (NDRTs) [2,7,8]. Takeover quality also degrades under high cognitive load [9], fatigue [10], passive fatigue arising from monotonous monitoring [11,12], vigilance decrement [13], reduced situational awareness [14,15], and over-trust or complacency [16–18].

The dominant response to these findings has been to improve the TOR itself: earlier alerts, richer multimodal signals, and better timing heuristics [19,20]. This line of work has produced genuine improvements, yet it addresses only the terminal moment of a longer re-engagement process [2,3,21]. It does not by itself determine whether the driver was in a state that is compatible with safe re-engagement before the alert was issued, whether the handover strategy should change under vehicle degradation or road-context urgency, or how a distracted, fatigued, or cognitively overloaded driver could be gradually brought back toward readiness before formal control transfer begins [2,3,21–23]. In particular, the challenge of passive fatigue requires proactive cognitive re-engagement rather than only a more salient terminal warning [11,12]. Transition of control should therefore be treated not only as a notification problem but as a cooperative reorientation problem [18,24].

The emergence of LLM-based in-vehicle conversational and agentic assistants adds a new layer to this already complex transition problem. Such systems can support natural contextual communication with the driver, for example by explaining why automation capability is degrading, by prompting soft vigilance before a formal TOR, or by adapting the interaction register to the current transition phase. Their value depends on the architecture in which they are embedded: in a safety-relevant takeover context, conversational assistance must be grounded in verified semantic state, bounded by explicit human-factors and safety policies, supervised at runtime, and channeled through validated HMI outputs.

In this paper, the term companion AI refers to an assumed LLM-based in-vehicle conversational or agentic assistant used as an advisory interaction component during takeover support. The contribution is defined at the architectural level: it investigates how such a component can be embedded within a digital-twin-mediated human–AI interaction framework so that its outputs are grounded in structured driver, vehicle, and context state, constrained by human-factor-derived guardrails, validated by runtime governance, and delivered through the HMI within a strict advisory/control boundary. Within this framework, trustworthy companion AI (TCAI) denotes the architectural trust layer that makes the companion AI component suitable for bounded takeover-support interaction.

Building on this framing, we model transition of control as a human-aware cooperative process that begins before any formal TOR and extends through driver reintegration into the driving task. Human, vehicle, and context/road digital twins provide structured semantic state abstractions through a semantic state interface enriched with confidence, freshness, provenance, and consistency metadata. The companion AI component proposes advisory interaction actions over these abstractions, while the TCAI layer governs those proposals

through bounded reasoning, policy enforcement, runtime watchdog validation, trust-dynamics modeling, and deterministic fallback. Safety-critical vehicle-control and MRC functions remain assigned to the deterministic vehicle-control stack, while the authorized output path of the TCAI layer is validated multimodal HMI delivery.

Figure 1 illustrates representative interaction situations in which such a companion AI component may operate. These situations are not intended as a fixed sequential workflow but as examples of the kinds of support that may be required around transition of control: maintaining driver engagement during automated operation, providing vigilance support when passive fatigue or disengagement is detected, explaining automation limitations to support trust calibration, and delivering concise directive guidance when manual takeover becomes necessary. The figure also highlights the architectural boundary that is central to the proposed approach: the companion AI component provides advisory information, explanations, and prompts through validated HMI outputs, while safety-critical vehicle-control and MRC functions remain assigned to the deterministic control stack.

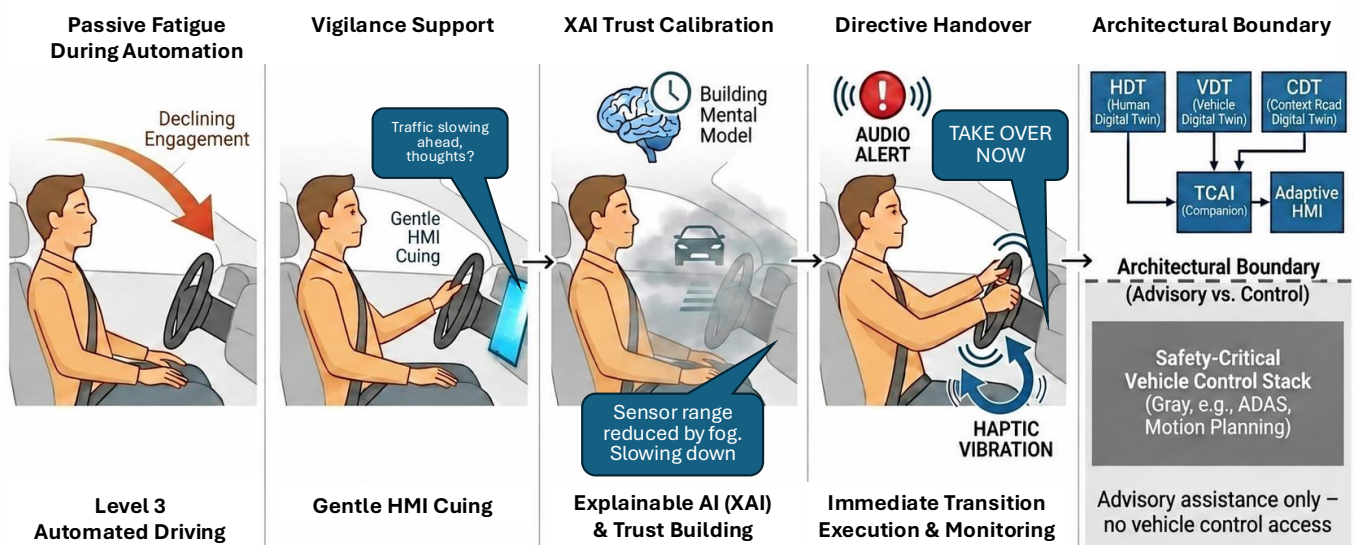


Figure 1. Illustrative interaction situations for digital-twin-mediated human–AI takeover support, including vigilance support, XAI-mediated trust calibration, directive handover guidance, and the advisory architectural boundary separating the companion AI component from safety-critical vehicle control. Limited graphical elements of this illustrative figure were prepared with support from FigureLabs.ai.

The novelty therefore lies in framing takeover support as a coherent system-level challenge requiring the integration of driver-state awareness, vehicle-degradation awareness, road-context awareness, bounded AI-mediated interaction, runtime governance, and deterministic fallback. The current work contains many of these elements separately, but the architectural choices required to combine them while preserving advisory/control separation remain underdeveloped.

Contributions

The main contributions of this paper are as follows:

1. We provide a narrative synthesis of fragmented approaches to takeover support, identifying persistent gaps across driver monitoring, adaptive HMI, trust calibration, explainability, and conversational assistance, and argue that general-purpose conversational or generative AI assistants change the architectural problem by requiring state grounding, bounded authority, and runtime governance.

2. We compare the proposed framework with classical human assistance systems (HASs) and cognitive assistance approaches, explaining the different architectural problem addressed here: digital-twin-mediated advisory-only human–AI interaction under uncertainty rather than vehicle-control or safety-management automation.
3. We consider takeover as a human-aware cooperative process, distinguishing it from the conventional framing of takeover as a one-shot TOR event and proposing system-level support for *soft vigilance* [12] as an intermediate operational target for safer re-engagement.
4. We propose a conceptual architecture in which human, vehicle, and context/road digital twins provide structured semantic state abstractions through a semantic state interface enriched with confidence, freshness, provenance, and consistency meta-data, enabling the TCAI layer to govern companion AI interaction proposals before HMI delivery.
5. We describe a candidate hybrid technical architecture for the TCAI layer and the surrounding digital-twin-mediated framework, combining probabilistic temporal models for state estimation, formally specified policy logic for phase selection, controlled generation by the companion AI component, and a deterministic watchdog for validation and fallback.
6. We present a technical development roadmap addressing driver-state estimation, latent driver mental-model representation, trust calibration, adaptive policy selection, explanation generation, runtime governance, and staged validation.
7. We make the advisory/control boundary explicit by assigning companion AI interaction to the validated HMI-delivery path and by keeping safety-critical vehicle-control and MRC functions within the deterministic vehicle-control stack.
8. We discuss ethical and interaction-design risks inherent to companion-mediated takeover support, with particular attention to over-reliance, false reassurance, anthropomorphic framing, calibrated trust, and driver veto or partial-disabling mechanisms.

The remainder of the paper is organized as follows. Section 3 discusses why current takeover-support approaches remain fragmented across driver monitoring, adaptive HMI, trust calibration, explainability, and conversational assistance. Section 4 introduces the proposed human-aware cooperative takeover-support vision and its core principles. Section 5 presents the digital-twin-mediated architecture and the TCAI layer. Section 6 describes runtime governance, degradation-aware operation, veto handling, and deterministic fallback. Section 7 outlines the validation agenda, while Section 8 presents the technical development roadmap. Section 9 discusses ethical and interaction-design considerations, and Section 10 concludes the paper.

2. Materials and Methods

Use of Generative AI Tools for Graphical Material

FigureLabs.ai (<https://figurelabs.ai/>, accessed on 26 May 2026) was used only to support the preparation of limited graphical elements in Figure 1, including illustrative visual components related to the driver silhouette, dashboard, and fog scenario. The conceptual design, scientific meaning, caption, and final validation of the figure were carried out by the authors. No generative AI tool was used to generate original research data, perform empirical analysis, determine study design, or interpret data.

3. Why Current Takeover Support Is Still Inadequate

The extensive TOR literature provides an essential empirical foundation, but its limitations as a basis for system-level transition design have become increasingly apparent [2,25]. The existing research has generated valuable knowledge about alert timing [2,26], modal-

ity [8], driver state estimation [27,28], trust [29–31], explanation [32,33], and conversational interaction [12,34,35]. What appears still underdeveloped is a unifying transition-of-control strategy that can integrate these capabilities into a coherent and bounded runtime assistance process. We organize our critique across five recurring dimensions involved in the TOR discussion: driver monitoring in isolation, adaptive HMI without unified state models, trust calibration without system-level orchestration, explainability disconnected from transition strategy, and conversational assistants without safety-relevant governance.

3.1. TOR Research: Accomplishments and Limits

Research on TORs has clarified several empirically robust phenomena. Takeover time depends strongly on the lead time provided, driver engagement in secondary tasks, and the complexity of the traffic situation at the moment of the request [7,9,19].

A recent systematic review and meta-analysis shows that takeover lead time and traffic volume are significant predictors of takeover time [2]. Other studies further show that engagement in non-driving-related tasks, particularly manual interactions such as smartphone use, substantially increases reaction times [4], while environmental complexity increases the cognitive effort required for drivers to re-enter the driving loop [23]. Multimodal alerts combining auditory and vibrotactile signals generally outperform unimodal presentations [4,8,20]. The cited studies consistently show that auditory–tactile takeover cues reduce response times compared to purely visual or unimodal alerts, partly because they bypass the visual channel, which is often already overloaded by non-driving-related tasks (NDRTs). Driver age, fatigue, and habituation further modulate performance [10,36,37]. These contributions are important, but TOR research remains predominantly reactive in orientation. The alert is still treated as the decisive event rather than as the endpoint of a managed preparatory process. There is only limited work on how the system should *prepare* the driver in the seconds or minutes before an anticipated ODD boundary, or on how monitoring, context assessment, interaction planning, and trust-sensitive communication should be coordinated as phases of a single transition strategy [3,25,38]. A driver who has been deeply disengaged from the driving task for around 15 min cannot simply be brought back to full readiness by an alert [12,14,15]. What is largely absent is a model of transition support that includes graduated reorientation, intermediate targets such as soft vigilance [12], and escalation policies adapted to both driver state and situational urgency [21,24,25]. The soft-vigilance state, as used in this architecture, requires explicit theoretical grounding distinct from adjacent constructs in the driver-state and situational awareness research. We define it operationally as a pre-cognitive reorientation state in which the driver: (a) directs perceptual attention toward the driving environment without sustained active monitoring; (b) maintains autonomic activity within a non-stress range, as indexed by heart-rate variability (HRV/RMSSD) and tonic electrodermal activity; and (c) exhibits gaze behavior consistent with environmental scanning rather than secondary-task engagement, as reflected in gaze dispersion above a task-disengagement threshold and an elevated percentage of eyes-on-road time [20,39]. This definition draws on Mackworth’s foundational characterization of vigilance as a readiness to perceive and respond that does not presuppose focused attentional engagement [40], on the psychophysiological operationalization of driver vigilance and its measurement across manual and automated driving contexts reviewed by Cabrall et al. [11], and on the well-established arousal–performance relationship—first formalized as an inverted-U function by Hebb [41] and recently confirmed in perceptual decision-making tasks by Nieuwenhuis [42], which predicts that an intermediate level of arousal, rather than maximal alertness, optimizes sensitivity to incoming environmental stimuli. No single proxy is individually sufficient; soft vigilance is estimated from the convergent pattern across gaze, autonomic, and behavioral channels,

with each contributing weighted evidence to the HDT state estimate. Soft vigilance is explicitly distinguished from Endsley's theory of situation awareness [43], with which it shares surface similarity. SA Level 1 in Endsley's framework denotes the active perception of relevant environmental elements and represents an already-achieved cognitive outcome; soft vigilance, by contrast, is the pre-cognitive enabling condition from which SA Level 1 may begin to be reconstructed. A driver in soft vigilance has reoriented perceptual resources toward the driving environment but has not yet integrated that information into a coherent situational model.

3.2. Driver Monitoring Systems in Isolation

Driver monitoring systems (DMSs) have advanced considerably, with modern approaches estimating drowsiness, distraction, cognitive load, and emotional state from camera-based gaze tracking, physiological sensors, and behavioral indicators [27,39,44,45]. UN Regulation 157 [46] now requires DMS integration in automated lane-keeping systems, reflecting a global regulatory trend. The critical limitation of contemporary DMSs is not only sensing accuracy but architectural role. These systems produce driver-state estimates that are largely consumed internally for binary warnings or automation gating without being exposed as structured machine-readable state representations to higher-level reasoning systems. Ref. [22] criticizes the binary nature of current automation-support systems, which tend to communicate system capability in overly discrete terms (e.g., available vs. unavailable) without conveying more nuanced information about operational uncertainty or performance limitations. Estimates are often coarsely discretized, lack calibrated uncertainty quantification, and do not distinguish sufficiently between transient attentional lapses, persistent cognitive unavailability, and intermediate readiness conditions. When DMS outputs are used for transition-related decisions, this is typically done through simple threshold logic rather than through a principled state-space model of driver readiness that could inform a phased and graduated transition strategy [21,47]. Ref. [3] observes that current terminology and transition-support systems still lack a sufficiently precise definition of what constitutes an "appropriate readiness state", making it difficult to distinguish between physical availability and actual cognitive readiness for takeover. In other words, the limitation is not simply that DMS observes the driver imperfectly but that it rarely exposes semantically reusable readiness information to a broader transition-management layer. As [24] advocates, a human-centered AI (HCAI) approach should aim to move beyond isolated assistance components and toward a more integrated joint cognitive system in which information is coordinated at a system-wide level.

3.3. Adaptive HMI Without Unified State Models

The adaptive HMI literature has explored modality switching, haptic alerts, ambient lighting cues, and dynamic information presentation for driving contexts [4,8,36,48]. Some work addresses driver-state responsiveness in interface design [49–52]. However, most adaptive HMI studies optimize interface behavior for isolated variables, such as the driver's visual attention state or the urgency of the driving situation, without access to an integrated model of driver state, vehicle status, and road context simultaneously. Transitions of control require exactly this multi-dimensional awareness: the same alert delivered to a well-rested attentive driver in low-complexity traffic requires a fundamentally different interaction design than an alert delivered to a fatigued driver approaching a complex interchange in adverse weather [35]. Without a unified state model from which the interface draws structured confidence-annotated information, HMI adaptation remains a collection of heuristics applied to partial information. The problem is therefore not the absence of adaptivity but the absence of *multi-factor coordinated adaptation*. An interface may be locally well-tuned

to a single monitored variable while remaining globally inappropriate for the transition situation as a whole. This becomes especially problematic when the interaction should evolve across phases, for example from low-intrusion support to contextual reorientation to directive prompting as urgency increases [3,12,21].

3.4. Trust Calibration Without System-Level Orchestration

Trust in automation is a well-studied construct, with established models distinguishing dispositional, situational, and learned trust, and with evidence that both over-trust (complacency) and under-trust (disuse) degrade system safety [53–55]. Driver-specific trust evolution trajectories are shaped by automation reliability history, transparency, and experience [56]. Experimental findings by Lee et al. [30] further show that continued interaction with the system, including exposure to limitations and failures, dynamically influences trust over time, while Lee and Pitts [57] highlight the importance of first impressions and information consistency in early trust formation. Zhang et al. [55] additionally emphasize that learned trust progressively emerges through experience-based learning processes grounded in prior interactions with the automation system. Calibrated trust, where driver reliance matches actual automation reliability [38], is recognized as a key safety objective [17,31,34]. Research on trust calibration in driving contexts has, however, developed largely in isolation from the transition-of-control problem [38]. Trust measurement is rarely integrated into the real-time system architecture [25]. Trust-calibration interventions, such as automated reliability disclosures or adaptive transparency mechanisms, are often proposed and evaluated without reference to how they would be orchestrated alongside DMS outputs, context information, vehicle capability, and adaptive HMI during an actual transition sequence. Yet trust calibration is not merely a background process: during a transition, trust directly shapes whether the driver interprets the system's request as credible, timely, and actionable [4,23,31]. To the best of our knowledge, the current literature does not yet offer a published architecture that treats trust calibration as an explicit component of a unified runtime transition-of-control strategy [24,25].

3.5. Explainability Disconnected from Transition Strategy

Explainable AI (XAI) for autonomous driving has received substantial attention, addressing explanations for path planning decisions, object detection reasoning, and system failure modes [58–60]. Work in this area has rightly emphasized the distinction between post hoc and intrinsic explainability and between explanations aimed at developers versus end users [32,61,62].

What is largely absent is a treatment of explanations as *functional components of a transition-of-control strategy* [2,25]. When a Level 3 system requests a handover because it has detected an ODD boundary or a sensor degradation event, the explanation provided to the driver is not merely informative: it is part of the reorientation process. Tan and Zhang [21] support this perspective through a multi-stage takeover framework employing semantic voice guidance to facilitate gradual situational reorientation during transitions, while Gerber et al. [3] report expert consensus highlighting the importance of detailed explanations during requests-to-intervene (RTIs) to help drivers understand both why intervention is required and where the danger originates. Whether the driver can quickly and correctly understand why control is being returned, what the system expects from them, and what the specific challenge of the current situation is directly affects transition quality [23,63]. This is particularly important if explanations are expected not only to justify a TOR but also to help move the driver from deep disengagement toward a more cognitively prepared state [34,35]. The existing XAI research does not yet treat explanation

generation as an integrated, context-sensitive, and driver-state-aware process embedded within a phased transition strategy [25,32,33,59].

3.6. Conversational Assistants Without Bounded Safety Context

Recent interest in large language models (LLMs) as conversational in-vehicle assistants has generated proposals for natural language interaction in driving contexts [12,35,64,65]. These systems can potentially provide richer and more adaptive verbal interaction during transitions, including contextual prompting, clarification, and reassurance. However, LLMs exhibit well-documented failure modes: hallucination, context window limitations, adversarial sensitivity, and the absence of any intrinsic grounding in safety-critical system states [66]. An LLM interacting with a driver during a safety-relevant transition without architectural bounds, structured state grounding, and runtime governance would introduce risks that are difficult to justify without additional safeguards [24,25,35]. Sourav et al. [25] explicitly identify hallucinations and the lack of transparency in LLM decision-making processes as major challenges, further noting that these systems still depend on external perception modules and must operate within the constraints of onboard hardware. Xu et al. [24] similarly discuss the risks of misleading or harmful content generation, the difficulty of aligning LLM behavior with human values and safety requirements, and the limitations of approaches focusing on isolated technical components without considering system-level coordination and the broader sociotechnical dimensions of safety. Markelius et al. [35] additionally observe that documented validation and verification (V&V) strategies for these systems in critical driving scenarios remain largely immature. The core issue is therefore not the use of conversational AI *per se* but its deployment without state grounding, bounded authority, and deterministic safeguards. Conversational assistance during transitions is a promising capability, particularly when interaction must evolve across different phases of re-engagement; however, it requires safety-aware constraints, structured semantic inputs, and fallback separation, which remain insufficiently developed in current transition-support architectures.

3.7. The Fragmentation Problem

Taken together, these five bodies of research address the transition-of-control problem in partial non-integrated ways. Driver monitoring systems observe the driver but typically do not expose semantically rich readiness states to a higher-level transition manager. Trust-calibration approaches shape communication but are not orchestrated with vehicle degradation, context urgency, and interface escalation. Adaptive HMI systems optimize presentation but often lack a unified model of driver, system, and road state. Explainability systems provide reasons but not as phase-aware elements of a transition strategy. Conversational systems interact but without strong safety governance. The main gap emerging from this research is not the lack of sensing modules or interface modalities *per se* but the limited integration of these elements into a unifying transition strategy that can coordinate them as parts of a bounded, runtime assistance architecture. This is the gap that the present vision addresses. Table 1 summarizes the main fragmentation patterns emerging from the literature discussed above.

The proposed digital-twin-mediated framework is intended to address this integration gap by combining semantically structured state representations, companion AI reasoning, adaptive interaction, runtime governance, and deterministic fallback within a single takeover-support architecture.

Table 1. Main fragmentation patterns in transition-of-control support research.

Literature Strand	Typical Focus	Missing for Unified Transition Support
TOR and takeover studies [2,21]	Lead time; alert modality; traffic complexity; NDRTs	Reactive framing; limited pre-transition support; no phased orchestration
Driver monitoring systems [27,47]	Distraction; drowsiness; fatigue; cognitive load	No semantic state export; limited uncertainty handling; weak integration with transition reasoning
Adaptive HMI [8,49]	Modality adaptation; salience; urgency-responsive presentation	Partial-state adaptation; no unified driver–vehicle–context model
Trust calibration & XAI [34,38]	Reliance; transparency; explanation; mental-model alignment	Not orchestrated as explicit components of phased transition support
Conversational assistants/LLM-based support [12,25]	Natural language; contextual prompting; clarification	No structured state grounding; no bounded authority; limited runtime governance
Overall gap	Partial solutions; isolated capabilities	Missing unified bounded architecture for state-aware, explanation-aware, and degradation-aware transition support

3.8. Relation to Classical Human Assistance Systems

The proposed framework is related to, but not a replacement for, classical human assistance systems (HASs) and cognitive assistance. Adaptive assistance, driver monitoring, and state-aware support are well established. Lin et al. [67] define HASs as systems that support human agents by monitoring operator state, task context, and system status, and by providing assistance through warnings, decision support, or automation adaptation. Cai and Lin [68] similarly frame cognitive assistance as a means of maintaining human engagement in human–machine collaboration. In the automotive domain, driver monitoring, fatigue recognition, and driver behavior modeling systems have long used probabilistic and machine learning methods, including dynamic Bayesian networks [69] and artificial neural network models of driver handling behavior in driver–vehicle–environment systems [70], to infer driver state, workload, or behavior and to support warnings or automation gating.

The distinction is therefore not that earlier systems lacked adaptive or intelligent support. The difference is the architectural problem considered here. Classical HASs typically concern assistance within a defined human–machine system and may be integrated with vehicle safety-management functions. The present framework asks how an assumed LLM-based companion AI component can interact with a driver in a safety-relevant takeover context without becoming ungrounded, unbounded, or safety-critical. This requires structured digital-twin state mediation, explicit advisory/control separation, runtime governance of generated or selected outputs, and deterministic fallback that is independent of the companion AI component.

The comparison in Table 2 is therefore intended as a scope distinction, not as a claim of superiority. Classical HASs remain foundational for driver-state monitoring and adaptive assistance; the present framework defines a higher-level interaction architecture for bounded human–AI takeover support under uncertainty.

In summary, the framework is not a remnant of HASs and does not dismiss classical HASs. It addresses the complementary problem of structuring, bounding, and governing human–AI interaction when an assumed LLM-based companion AI component may generate context-sensitive language during takeover support. The relevant difference is not

which specific LLM implementation is used but how interaction among AI, the driver, digital twins, HMI, governance, and safety-critical systems is mediated.

Table 2. Scope distinction between classical HASs and the proposed digital-twin-mediated human–AI takeover-support framework.

Dimension	Classical HAS / Cognitive Assistance	Digital-Twin-Mediated Framework
Architectural role	May be embedded in vehicle control or safety-management functions; may gate automation modes or trigger alerts [67,68]	Advisory-only interaction framework; companion AI component remains separated from vehicle control and MRC stack; no write-access to actuation [24,25]
State representation	Internal signals; binary or threshold-based; consumed locally within decision logic [67,69]	Human, vehicle, and context/road digital twins expose semantic state through an interface with confidence, freshness, and provenance metadata [25,47]
Human modeling	Snapshot estimates of readiness, drowsiness, and distraction for immediate safety management [27,69]	Dynamic HDT with trust profile and latent driver mental-model estimate inferred from interaction history [47,56]
Interaction mechanism	Discrete alert/warning modalities; urgency-triggered modal switching [8,48]	Multi-phase adaptive interaction across monitoring, anticipatory preparation, and transition execution; XAI-mediated trust calibration [12,21,34]
Safety boundary	Often safety-relevant or connected to safety-management logic depending on implementation [67,69]	Companion AI component is non-safety-critical and advisory; vehicle-control and MRC safety remain with the deterministic control stack [24,25]
Runtime governance	Typically managed through thresholds, validation, and functional-safety processes [67,68]	Dedicated watchdog with formally specified policies; companion AI outputs validated before HMI delivery; non-learned deterministic fallback independent of companion AI reasoning [24,25]
AI technology scope	Narrow trainable models, such as DBNs, SVMs, CNNs, or rule-based components [69]	Assumed LLM-based conversational or agentic companion AI component used for interaction, creating hallucination, boundary-negotiation, and false-reassurance risks that require TCAI governance [35,64–66]

4. Human-Aware Cooperative Takeover Support: Vision and Principles

4.1. A Motivating Scenario: Level 3 Transition Under Reduced Visibility

To ground the principles articulated in this section, consider the following illustrative scenario. A driver is traveling at motorway speed on a monotonous extra-urban road with SAE Level 3 automation engaged. After approximately twenty minutes of automated operation, the HDT registers extended fixation durations and attenuated heart-rate variability consistent with the onset of *passive fatigue*, namely the progressive cognitive disengagement that occurs when a driver is relieved of active vehicle control over an extended period [10,12]. The driver has not been prompted to monitor the road and has partly withdrawn attention from the driving environment.

Simultaneously, the context/road digital twin detects a fog bank ahead. Its trajectory model predicts that visibility will fall below the automation system’s reliable perception threshold within approximately ninety seconds, placing the vehicle on a path toward an ODD boundary. The vehicle digital twin begins reporting declining LiDAR confidence values consistent with the deteriorating atmospheric conditions. A conventional architecture at this point would wait for the ODD boundary to be reached, then issue a generic takeover alert applied to a driver whose readiness state has not been managed.

Within the digital-twin-mediated framework, the companion AI component, operating under the TCAI layer, uses integrated HDT, VDT, and context state to identify the developing scenario well before the ODD boundary is reached. Recognizing that the driver readiness estimate is below the threshold that is considered to be adequate for an unassisted transition, the system initiates the first phase of graduated pre-transition support: a light

contextual prompt calibrated to redirect attention toward the driving environment without provoking acute alarm. The aim of this first-phase interaction is to guide the driver toward *soft vigilance* [12]: a cognitive state in which the driver remains physically relaxed but is reoriented toward the environment and therefore better positioned to respond when a formal transition is required. Soft vigilance is not equivalent to sustained high-alertness monitoring, which is cognitively costly and difficult to sustain over the durations typical of Level 3 operation; it is a more modest and operationally achievable intermediate target.

As fog density increases and the VDT's perception confidence crosses a pre-configured warning threshold, the companion AI component, operating under the TCAI layer, transitions to a second communication phase. Rather than emitting an undifferentiated alarm, it delivers a concise contextualized explanation: it identifies the cause of the system's deceleration as reduced sensor range, states the anticipated timeline to the handover request, and briefly characterizes the driving challenge ahead. This XAI-mediated trust calibration allows the driver to update their mental model of the system's operational state accurately, preventing both the complacency induced by an over-trusting misbelief that the system will manage the fog and the disproportionate alarm that an unexplained alert would produce [34]. Contextual explanations that tie system behavior directly to observable environmental features (fog reducing laser range and ice impairing ABS responsiveness) have been shown to increase transparency perception and support more stable trust alignment than generic capability notifications [34,59].

When fog density crosses the ODD exit threshold, the companion AI component, operating under the TCAI layer, shifts to a third communication phase. The explanatory contextualizing register gives way to brief directive action-oriented guidance. This transition in communication style reflects the empirically grounded principle that working-memory load during an imminent manual takeover demands high-priority low-latency signals rather than the elaborated framing appropriate for preparatory phases [43,71].

The driver, having been progressively reoriented rather than abruptly interrupted from a state of passive fatigue, would be expected to resume control from a more favorable situational-awareness condition than a driver who encountered an unmediated alert sequence. At the same time, this should not be read as assuming perfect cooperation or guaranteed takeover quality: soft vigilance is an enabling intermediate condition, not a guarantee of successful manual control under all urgency levels.

This scenario, condensed here as a conceptual illustration, not as a report of completed empirical work, motivates several design requirements for any cooperative transition-support architecture: a mechanism to detect and manage passive fatigue; a defined intermediate target state (soft vigilance) short of full situational awareness; XAI-based trust calibration grounded in observable context; phase-sensitive communication that adapts register to urgency; and a stress-aware interaction strategy that avoids unnecessarily alarming prompts in the pre-transition window.

4.2. Reframing the Problem

Building on emerging perspectives in human factors and human-centered automation research [18,21,24,25], we frame takeover support as a form of *human-aware cooperative transition of control* (HACTC). Within this framing, transition is treated not as an isolated event but as a process; not as a one-directional command but as a cooperative exchange; and not as a fixed alerting protocol but as a human-in-the-loop challenge requiring dynamic adaptation across evolving driver, system, and contextual conditions.

In this framing, transition of control encompasses:

- **Pre-transition awareness:** continuous monitoring of driver state, vehicle status, and road context to identify developing transition preconditions.

- **Graduated pre-transition prompting:** proactive graded assistance to reorient the driver toward the driving task before a formal TOR is necessary when preparation time permits.
- **Context-adaptive handover orchestration:** dynamic adjustment of the transition sequence based on the urgency of the triggering event, the complexity of the road context, and the momentary driver-readiness state.
- **Situational reorientation:** active provision of context-sensitive explanations allowing the driver to understand rapidly why control is being returned and what the immediate driving challenge is.
- **Post-handover support:** monitoring of driver re-engagement quality following the transition, with continued assistance if reintegration is incomplete.

This perspective treats the TOR not as the transition itself but as only one interaction point within a broader multi-stage re-engagement process through which the driver progressively reconstructs readiness, contextual understanding, and control capability.

4.3. Relevance to Level 4 Fallback Conditions

While this framing is motivated primarily by the structural challenge of Level 3 automation, it is also conceptually relevant to degraded and fallback-related Level 4 conditions. SAE Level 4 systems are designed to operate without human intervention within their ODD, but ODD boundaries are not always sharply defined in practice, and unexpected degradation events can produce de facto transition scenarios even in systems nominally certified for Level 4 operation [15,72]. When Level 4 automation degrades due to sensor failure, unexpected ODD exit, or operational anomalies, the driver may face a transition whose suddenness and unexpectedness exceed the demands of a planned Level 3 handover.

A TCAI-governed companion AI component capable of detecting degradation early and managing pre-transition support in this context would therefore provide safety value beyond the canonical planned-transition scenario. The proposed vision is not intended as a generic framework for all Level 4 HMI design; rather, it is a framework for managing the boundary between human and automated driving agency when that boundary is crossed under conditions of uncertainty, degradation, or reduced preparation time.

4.4. Core Principles

The proposed vision is organized around six core principles:

P1—Assistance, not control. The companion AI component provides advisory, explanatory, and interactional support. It does not and must not issue or modify vehicle-control commands. The architectural separation between the advisory layer and the safety-critical control stack is a core design requirement.

P2—Semantic state reasoning, not raw-sensor reasoning. The companion AI component does not reason directly over sensor feeds, LiDAR point clouds, or camera images. It reasons over structured human-interpretable state representations provided by digital-twin abstractions, accompanied by confidence scores and data-freshness metadata. This abstraction layer provides robustness, interpretability, and a clean architectural interface.

P3—Graduated assistance targeting soft vigilance. Transition support is graduated in intensity as a function of driver unreadiness, transition urgency, and situational complexity. A rested attentive driver approaching a routine ODD boundary in benign traffic receives minimal non-disruptive support. A fatigued cognitively loaded driver approaching a complex transition in adverse conditions receives progressively earlier, more salient, and more explanatory engagement. A key intermediate goal of pre-transition support is to guide the driver toward a state of *soft vigilance* [12]: cognitively re-engaged with the driving environment yet not acutely alarmed. Soft vigilance is a more achievable and less cognitively costly

target than sustained high-intensity monitoring, and it provides a realistic preparatory condition from which effective manual control may be resumed.

P4—Contextualized explainability for reorientation and trust calibration. TCAI-mediated companion AI outputs that reach the driver should be accompanied, where useful and permissible under urgency constraints, by concise driver-appropriate explanations grounded in the current state of the vehicle and road. Explanations are not generic system notifications; they are situationally specific, calibrated to the driver’s inferred understanding, and structured to support reorientation and more appropriate trust calibration [59].

P5—Phase-sensitive communication and stress regulation. The interaction style of the companion AI component should not be fixed but should evolve across operationally distinct phases of the transition process. During the *monitoring and maintenance* phase, communication is light and contextual, aimed at sustaining soft vigilance without imposing unnecessary cognitive load. During the *anticipatory preparation* phase, communication shifts to contextualized XAI explanations that calibrate the driver’s mental model to the system’s current operational state and anticipated capability limits [34]. During the *transition execution* phase, communication becomes brief, direct, and action-oriented, avoiding the verbal elaboration that would impair working-memory performance under time pressure [43,71]. This phase sensitivity is not merely a stylistic preference; it is a functional design requirement because both cognitive demands and acceptable stress levels vary substantially across phases.

P6—Trustworthiness by design. The TCAI layer makes the companion AI component trustworthy only insofar as it is bounded, governed, transparent, and supervised. The companion AI component’s action space is explicitly constrained, its reasoning is logged, uncertainty is represented through the semantic state interface, and fallback to deterministic behavior remains available under degradation.

5. Digital-Twin-Mediated Human–AI Interaction Architecture

5.1. Architectural Overview

The proposed digital-twin-mediated architecture is organized around five architectural blocks: (i) a digital-twin layer comprising three coordinated modules: the human digital twin (HDT), the vehicle digital twin (VDT), and the context/road digital twin (CDT); (ii) a semantic state interface; (iii) a companion AI component; (iv) an adaptive HMI subsystem; and (v) the TCAI layer, which includes bounded reasoning, policy enforcement, runtime governance, watchdog supervision, fallback, and advisory/control separation. Figure 2 provides a conceptual view.

The architecture can be read from left to right. Multimodal sensing inputs (including camera, LiDAR, radar, V2X, and physiological or behavioral signals where available) feed the three digital-twin modules. The HDT tracks driver-related variables such as readiness, cognitive load, stress, and soft-vigilance level; the VDT tracks automation mode, ODD proximity, sensor health, and perception confidence; and the CDT tracks road geometry, traffic, weather, and urgency conditions. These twins do not expose raw sensor streams to the companion. Instead, they provide structured state estimates to the semantic state interface, which enriches them with confidence scores, freshness timestamps, and provenance information. The companion AI component proposes interaction strategies over these semantic abstractions, ranging from soft-vigilance maintenance to contextual explanation and trust calibration to brief directive guidance under time pressure. The resulting interaction policy is realized through the adaptive HMI only after the TCAI layer’s policy-enforcement and runtime watchdog mechanisms validate companion AI outputs before HMI delivery. A hard architectural boundary separates the entire advisory assistance layer from the safety-critical vehicle-control and minimum risk condition (MRC) stack. The

companion has no write-access to actuation. Edge-first processing and privacy-preserving inference are assumed, while cloud connectivity is treated as optional augmentation rather than as a safety-critical dependency.

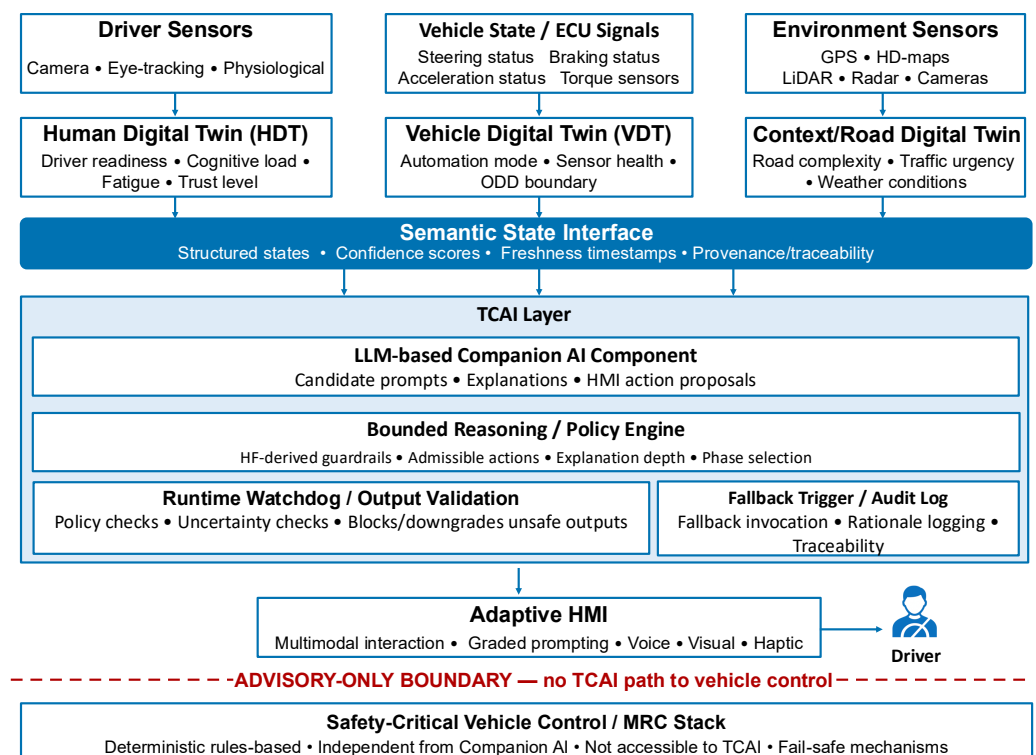


Figure 2. Conceptual architecture of the proposed digital-twin-mediated human–AI interaction framework for takeover support. Human, vehicle, and context/road digital twins expose structured semantic states through a semantic state interface enriched with confidence, freshness, and provenance/traceability metadata. The TCAI layer embeds the LLM-based companion AI component within bounded reasoning, policy enforcement, runtime watchdog/output validation, and fallback/audit mechanisms. Only validated interaction proposals reach the driver through the adaptive HMI, while the advisory-only boundary separates the TCAI layer from the safety-critical vehicle-control and minimum risk condition (MRC) stack.

The architecture is motivated by three design imperatives. First, the companion AI component should reason over structured semantically interpretable representations rather than raw sensor data because only structured representations can carry the confidence, freshness, and provenance metadata required for degradation-aware reasoning. Second, the companion AI component must be isolated from the vehicle’s actuation interfaces because its inherent uncertainty and adaptiveness make it unsuitable as a safety-critical component. Third, all outputs of the companion AI component must pass through the TCAI layer’s deterministic runtime watchdog before HMI delivery because even a well-designed advisory AI may produce inappropriate outputs under out-of-distribution or degraded conditions. The purpose of the architecture is therefore not merely to integrate modules but to support transition timing, communication strategy, and escalation policy under uncertainty.

Recent work on mobility and transportation digital twins has shown that DTs can support dynamic representations of humans, vehicles, traffic, and infrastructure by combining sensing, V2X communication, cloud–edge processing, simulation, and prediction [73,74]. In particular, mobility-DT frameworks explicitly organize the mobility system around human, vehicle, and traffic digital counterparts, while recent smart-mobility DT platforms

for connected and automated vehicles demonstrate the operational relevance of latency, reliability, and roadside/cloud integration [73,74]. In parallel, driver- and human-centered DT research has emphasized multimodal state fusion, personalization, time variance, and the need for clearer conceptual foundations when modeling human state and behavior [75–77]. The architecture proposed here builds on these developments but assigns DTs a more specific role: they act as semantically structured state providers for a bounded companion AI component that must decide how to support takeover interaction under uncertainty.

5.2. Human Digital Twin (HDT)

The human digital twin maintains a continuously updated model of the driver's operational state. This use of the HDT is consistent with recent work on driver digital twins and human digital twins, which emphasizes multimodal state fusion, personalized modeling, time variance, and the need to represent the human not as a static user profile but as a dynamic individual-specific entity [75,76]. Its outputs include, at minimum, driver readiness level, estimated cognitive load, fatigue index, momentary attention state (gaze direction and eyes-on-road status), emotional arousal, stress indicator (e.g., derived from heart-rate variability or electrodermal activity where available), a soft vigilance estimate reflecting the degree to which the driver is passively disengaged versus environmentally reoriented, and a trust-calibration profile encoding the driver's accumulated experience with the current automation system [12,47].

The soft-vigilance state is a key intermediate output of the HDT. It is distinguished from a binary alert/fatigued classification by capturing the continuum between passive fatigue, where the driver is cognitively withdrawn and physiologically relaxed but environmentally disengaged, and full active monitoring. The TCAI layer uses this estimate to determine whether pre-transition support should aim to restore soft vigilance as an intermediate target or escalation toward a formal transition protocol is required because adequate readiness cannot be established within the available preparation time.

The HDT draws on inputs from camera-based gaze tracking, physiological sensors where available, and inferred behavioral indicators. Crucially, the HDT does not expose raw sensor data to the TCAI layer. It exposes structured state variables with associated confidence intervals and data-freshness timestamps. A readiness estimate derived from a gaze tracker that has been occluded for the past three seconds carries a different confidence and freshness tag than one derived from a full physiological sensor suite with continuous signal quality. The TCAI layer is expected to reason appropriately over these metadata [53].

This sensing perspective is aligned with the human-centric DT literature, where motion sensors, biological sensors, computational intelligence, simulation, and human-facing interfaces are identified as key enabling technologies for constructing operational human-centered DTs [77].

The HDT also maintains a longitudinal trust profile capturing how the individual driver has historically responded to automation interventions, system-transparency events, and prior transitions [56]. This profile informs the selection of explanation styles and interaction modalities during the current transition.

This trust profile is used by the TCAI-layer Trust Dynamics Module described in Section 5.7.1.

5.3. Vehicle Digital Twin (VDT)

The vehicle digital twin maintains a continuously updated model of the automation system's operational status. Key state variables include current automation mode, ODD-boundary proximity (distance and estimated time to anticipated boundary), sensor health

per modality, actuator status, diagnostic fault codes, and the confidence of the underlying perception and planning stack [78,79].

The VDT does not relay raw sensor feeds to the TCAI layer. It exposes semantic status representations. A perception-confidence drop caused by temporary LiDAR occlusion is represented differently from one caused by a hardware fault, a calibration drift, or an internal diagnostic anomaly. This distinction matters for the companion AI component's reasoning: temporary occlusion may resolve within seconds, whereas a hardware fault or persistent internal degradation changes both transition urgency and the explanation appropriate for the driver. More generally, the VDT allows the companion AI component to reason not only about whether automation is degrading but about how that degradation affects the validity of the current assistance policy.

5.4. Context/Road Digital Twin

The context/road digital twin maintains a model of the driving environment relevant to the transition decision. Key state variables include road type and geometry, current traffic density and flow, weather and visibility conditions, proximity to known ODD boundaries (road-type transitions and geofenced zones), and the urgency level of the transition trigger. This twin draws on HD-map data, V2X communications, onboard environment perception, and historical traffic data.

The urgency state is particularly important. A transition trigger caused by an unanticipated vehicle cut-off in high-speed motorway traffic is qualitatively different from a planned transition at a geofenced urban boundary. Urgency, however, is not purely a property of the environment; it emerges from the interaction between environmental demands, current system capability, and the time available to prepare the driver. The CDT encodes this interaction in structured machine-readable form so that the companion AI component, operating under the TCAI layer, can calibrate both the urgency and the modality of its response.

This role is consistent with recent infrastructure-based DT work for cooperative, connected, and automated driving, where DTs are used to provide quality-controlled formalized machine-usable information collected and augmented from multiple sources to support automated-driving functions [80]. In the proposed architecture, this idea is specialized toward transition support: the context/road DT does not only describe the environment but contributes to estimating transition urgency, contextual complexity, and the admissible communication policy.

5.5. Semantic State Interface

The semantic state interface is the structured data contract between the three DT modules and the companion AI component governed by the TCAI layer. It is also the key mediation layer between deterministic state estimation and adaptive transition reasoning. For each state variable, the interface exposes:

- The **value** of the state variable (categorical, ordinal, or continuous as appropriate);
- A **confidence score** capturing the uncertainty of the estimate;
- A **freshness timestamp** and a configured staleness threshold beyond which the estimate should not be used for transition decisions without flagging;
- A **provenance tag** identifying which sensing modality or sub-model produced the estimate, supporting traceability.

This design is motivated by the need for robust operation under degradation. Without freshness and confidence metadata, the companion AI component may treat stale or uncertain state estimates as more reliable than they actually are. A companion AI component that receives only confidence scores without provenance cannot distinguish between a

degraded sensing chain, a transient observability problem, and an environmental condition that is inherently ambiguous. The semantic state interface makes these distinctions explicit and machine-processable. Its role is therefore not only to transport DT outputs but to normalize heterogeneous state estimates into a form that supports bounded reasoning, traceable explanations, and policy-compliant escalation under uncertainty.

Table 3 provides illustrative examples of semantic state variables that may be exposed to the companion AI component through the semantic state interface and governed by the TCAI layer.

Table 3. Illustrative semantic state variables exposed through the semantic state interface.

State Variable	Origin	Illustrative Role in TCAI-Governed Interaction
Driver readiness	HDT	Supports escalation timing and selection of anticipatory versus directive support
Soft vigilance level	HDT	Indicates whether pre-transition prompting should aim at reorientation before a formal handover request
Automation capability	VDT	Signals whether assistance should remain explanatory or shift toward urgent transition support under degradation
ODD boundary proximity	VDT/CDT	Supports transition urgency estimation and phase selection
Contextual urgency	CDT	Modulates interaction salience, explanation depth, and admissible response policy

5.6. State Consistency Management

The semantic state interface must manage consistency across asynchronous and heterogeneous digital-twin outputs. Because HDT, VDT, and CDT variables update at different rates, each semantic state should carry a timestamp, freshness threshold, confidence value, and provenance metadata. Timestamp alignment should use a bounded temporal window: variables outside their admissible freshness window should be marked stale or excluded from interaction-policy selection.

Conflict detection should operate at intra-twin and inter-twin levels. Intra-twin conflicts arise when sensing modalities or sub-models within one digital twin disagree, for example when gaze suggests road reorientation while physiology suggests passive disengagement. Inter-twin conflicts arise when driver, vehicle, and context states imply incompatible transition policies, for example low HDT readiness with high VDT/CDT urgency. Such conflicts should appear as consistency flags rather than be hidden inside a single fused score.

Arbitration should follow conservative policy-defined rules. When conflicts affect safety-relevant timing or explanation depth, the TCAI layer should preserve advisory boundedness and avoid overconfident assistance, for example by selecting the cautious readiness estimate, reducing explanation depth, downgrading output to a guarded template, or triggering deterministic fallback when the snapshot is too inconsistent for personalized assistance. All violations and arbitration decisions should be logged with timestamps, confidence scores, and provenance metadata for audit and policy refinement.

5.7. Companion AI Component and TCAI Layer

The companion AI component is the advisory reasoning and interaction component of the framework, while the TCAI layer is the architectural trust layer that grounds, bounds, and supervises it. The companion AI component operates on the structured states provided by the three digital twins and proposes candidate takeover-support interactions: what assistance to offer, in what modality, at what intensity level, and with what explanation.

The companion AI component, operating under the TCAI layer, is designed with the following explicit constraints:

- **Bounded action space:** The companion AI component can propose actions from a predefined set of interaction actions: generating verbal prompts, triggering visual displays, activating haptic signals, and composing situation-specific explanations. The TCAI layer's bounded reasoning and policy engine constrain the admissible action space, and the companion AI component cannot modify vehicle-control parameters.
- **Supervised operation:** All companion AI output proposals are logged, monitorable, and subject to runtime watchdog override before HMI delivery.
- **Explainable reasoning:** The companion AI component produces, alongside each interaction proposal, a structured rationale that can be inspected by the TCAI layer and, where appropriate, communicated to the driver.
- **Uncertainty-aware:** The companion AI component propagates and communicates uncertainty from the state interface through its reasoning to its output proposals. A transition-support action based on low-confidence driver-state estimates is flagged as such in the companion AI output log.

The architecture assumes an LLM-based companion AI component for conversational or agentic interaction, but it is intentionally agnostic about the specific commercial model, deployment stack, or prompting interface used for that component. The paper does not compare language-model architectures or propose training a new companion AI. Rule-based components, probabilistic graphical models, and symbolic or knowledge-based planners are instead treated as possible technologies for the surrounding digital-twin-mediated framework and TCAI layer, particularly for state estimation, bounded reasoning, policy enforcement, and verification-oriented governance [64,81–83].

5.7.1. Trust Dynamics Module

The TCAI layer should include an explicit Trust Dynamics Module that uses the HDT trust profile as a policy input and updates the trust estimate through interaction evidence. It does not infer a driver's internal beliefs directly; instead, it maintains a probabilistic longitudinal estimate of trust calibration from observable proxies: historical handover outcomes, compliance or non-compliance with prompts, driver feedback or clarification requests, exposure to automation limitations, transparency events, uncertainty communication, and post-transition performance. Trust is therefore treated as a dynamic latent state shaped by accumulated interaction history, not as a static user attribute [30,53,54,56,57].

The module should estimate whether current reliance is plausibly calibrated, over-trusting, or under-trusting relative to automation capability and context and should expose an uncertainty score for that estimate. Sparse interaction history or inconsistent evidence should therefore reduce the stability of the trust profile. Under TCAI governance, these outputs modulate companion AI interaction proposals: over-trust may call for limitation-focused explanations during anticipatory preparation, whereas under-trust may call for concise transparency cues that clarify current capability without false reassurance [34,38,84].

The Trust Dynamics Module also constrains interaction intensity. Apparent over-reliance should lead the TCAI layer to avoid reassuring or anthropomorphic language and to make uncertainty and system limits visible. Apparent under-reliance should not trigger persuasive reassurance but grounded explanations of current system state and fallback boundaries. In all cases, trust-dynamics estimates remain advisory policy inputs: they may influence explanation style, timing, and modality but cannot override urgency, state-consistency flags, watchdog vetoes, or the advisory/control boundary.

5.7.2. Hybrid TCAI-Layer Architecture: Candidate Technical Realization

Given the safety-relevant context, the assumed LLM-based companion AI component is not treated as the whole trustworthy interaction system. We therefore propose a *hybrid realization of the TCAI layer and surrounding digital-twin-mediated framework* with four components:

(i) Probabilistic temporal models for driver-state estimation. The HDT uses continuous uncertainty-quantified estimates of readiness, cognitive load, fatigue, soft vigilance, and arousal. Candidate approaches include DBNs [69], hidden Markov models (HMMs), or temporal neural models applied to multimodal physiological and behavioral time series. These models output calibrated state estimates and uncertainty, making their failure modes more characterizable than open-ended generative outputs.

(ii) Symbolic or rule-based logic for phase selection and admissible action policy. Transition-phase selection and admissible actions should be formally specified, for example through a decision table, finite-state machine, or constrained symbolic policy. This keeps escalation, explanation suppression, and fallback decisions outside learned or generative components.

(iii) Controlled natural-language generation for explanations. Explanation generation is associated with the assumed LLM-based companion AI component, but TCAI mediation should constrain it through guarded templates, retrieval over verified knowledge resources, or constrained prompting with output validation against the semantic state interface. These are governance strategies around the companion AI component, not alternative definitions of it.

(iv) Deterministic watchdog for output validation and fallback. The watchdog is non-learned and validates companion AI proposals against policy and state context before HMI delivery. If validation fails, fallback is activated; the watchdog must remain independently verifiable and implementation-independent from the component it supervises.

This decomposition supports modular development, validation, and auditing: each component has a defined function, characterizable failure mode, and relationship to governance. It is not a comparison of companion AI architectures; Section 8 elaborates how the surrounding framework components could be trained, specified, or calibrated.

5.8. Adaptive HMI Subsystem

The adaptive HMI translates TCAI-validated companion AI interaction proposals into concrete perceptible outputs delivered to the driver. It supports multimodal delivery: auditory speech output, visual dashboard and HUD cues, ambient lighting, and haptic actuation through the steering wheel or seat [20,48].

Table 4 summarizes the illustrative phase logic used by the companion AI component, under the TCAI layer, to adapt interaction objectives, modalities, and explanation style across different transition-support conditions.

Table 4. Illustrative phase logic for adaptive transition support.

Phase	Objective and Modality	Explanation Style
Monitoring/maintenance	Sustain or restore soft vigilance with low-intrusion ambient, subtle visual, or light verbal cues	Contextual and minimal
Anticipatory preparation	Reorient the driver and calibrate trust through verbal and visual guidance, with haptic preview where appropriate	Concise, situational, and causal
Transition execution	Support immediate control resumption through high-salience auditory and haptic guidance with minimal verbal content	Directive; elaboration deferred

The HMI adaptation logic is structured around the three communication phases introduced in Principle P5. During the **monitoring and maintenance phase**, the system operates with minimal intrusion: ambient cues and, where soft vigilance is estimated to be declining, light contextual conversational prompts designed to re-engage environmental attention without unnecessarily interrupting secondary activities. The communicative style in this phase is collaborative and contextually grounded rather than directive [12]. During the **anticipatory preparation phase**, the HMI shifts to explicit XAI-mediated trust calibration: concise verbal or visual explanations that link the system's observed behavior to identifiable environmental causes (e.g., reduced perception confidence attributed to fog density or sensor obstruction), allowing the driver to construct a more accurate mental model of the system's operational state [34,59]. During the **transition execution phase**, the system abandons elaboration entirely: haptic and high-salience auditory signals are prioritized, verbal content is reduced to brief action instructions, and any explanatory content is deferred to the post-transition window [43,71].

This phase structure also governs stress regulation. Research on stress in transition contexts has established that the driver's stress state is shaped not only by the transition event itself but also by the quality of preparation and the coherence of communication leading up to it. A driver who has received contextualized preparation and trust-calibrating explanation in the anticipatory phase experiences the transition event under substantially different cognitive and emotional conditions than one who receives only an unmediated alarm. The HMI must therefore be sensitive not only to current urgency but also to whether prior communication has plausibly brought the driver closer to the soft-vigilance target state.

Modality selection and intensity are governed by graduated protocols rather than fixed thresholds, supporting smooth escalation and de-escalation as transition conditions evolve. The companion AI component, under the TCAI layer, does not attempt to override driver discomfort through indefinitely increasing alert aggressiveness; instead, if a driver's stress indicators suggest that the current interaction policy is itself counterproductive, the system should escalate to the deterministic fallback protocol rather than compound arousal with poorly timed assistance.

Dynamic Explanation-Depth Policy

The TCAI layer should govern explanation depth dynamically rather than treating detail as a fixed HMI property. We conceptualize explanation depth as four levels: E0 cue-only output, E1 minimal directive, E2 concise causal explanation, and E3 contextual explanation including cause, expected evolution, system limitation, and driver-relevant implication. These are not quality ranks: higher depth is useful only when the driver has sufficient time and cognitive capacity [23,32,33,43,71].

Admissible explanation depth can be modeled as a policy function of transition urgency, estimated cognitive load, task complexity, and inferred mental-model mismatch. Urgency, cognitive load, and task complexity should reduce admissible depth when working-memory or perceptual demands are high [23,71]. Conversely, low urgency, moderate load, and persistent mental-model mismatch may permit E2 or E3 explanations for anticipatory trust calibration and situational reorientation [34,38].

A conservative rule follows: under imminent handover or high workload, the watchdog restricts the companion AI component to E0–E1; during anticipatory preparation with sufficient lead time, E2 may be permitted; and, during monitoring or post-transition phases, E3 may be permitted if grounded in fresh semantic state and unlikely to increase stress or over-trust. The runtime watchdog validates explanation depth with modality, salience,

and wording; if depth exceeds the admissible level, the output is downgraded according to the veto-handling logic in Section 6.3.

5.9. Architectural Boundary: Advisory vs. Safety-Critical

A core design requirement of the architecture is the strict separation between the advisory layer and the safety-critical vehicle-control stack. The TCAI layer has no write-access to vehicle-control parameters, no connection to actuation interfaces, and no ability to override or modify the behavior of the underlying automated-driving system or its minimum risk condition (MRC) logic.

This separation ensures that, even in the event of a companion AI malfunction, adversarial attack, or unanticipated reasoning failure, the safety-critical control behavior of the vehicle is unaffected. The vehicle’s MRC routine executes independently of the companion AI component’s status. The companion AI component is, by design, a non-load-bearing element of the vehicle’s safety architecture. This is not only a matter of architectural prudence but also of assurance: keeping advisory intelligence separate from deterministic control preserves a cleaner boundary for certification, runtime supervision, and safety-case argumentation. Table 5 summarizes the architectural permissions and restrictions.

Table 5. Architectural permissions and restrictions of the companion AI component.

Category	Permissions and Restrictions
Permitted	Read structured DT state exports; generate HMI interaction actions; produce rationale logs; communicate uncertainty to the driver
Restricted	Write to vehicle-control parameters; access actuation interfaces; modify MRC logic; suppress the runtime governance watchdog
Delegated	All safety-critical decisions remain with the vehicle’s deterministic control and MRC stack; all safety-relevant interventions pass through the runtime watchdog before HMI delivery

Security Boundary and Mandatory Access-Control Assumptions

The advisory/control boundary should be enforced through mandatory access-control policies and hardware/software isolation, not only design intent. In a deployable architecture, the companion AI component and TCAI-layer services should run in a sandboxed software domain, isolated process, container, virtualized partition, or dedicated non-safety-critical ECU domain separated from the safety-critical vehicle-control and MRC stack. The semantic state interface should expose only read-only policy-filtered state abstractions, and no TCAI-layer process should possess credentials, permissions, API endpoints, or bus-level access for writing vehicle-control parameters, issuing actuation commands, suppressing MRC execution, or modifying safety-critical controller configuration.

Mandatory access control should follow deny-by-default and least-privilege principles. The companion AI component may propose prompts, explanations, or HMI actions, but only to a controlled output queue supervised by the runtime watchdog. The only authorized outbound path from the TCAI layer is therefore validated HMI delivery; there is no authorized path to steering, braking, acceleration, powertrain, perception-stack configuration, or MRC execution. Policies and guardrails should be versioned, integrity-protected, and auditable.

This isolation assumption is part of the safety argument. If the companion AI component is compromised, unavailable, or unsafe, the watchdog can block HMI output and invoke deterministic fallback, but the component cannot alter vehicle control. Conversely, if the vehicle-control or MRC stack initiates a safety-critical maneuver, the TCAI layer may

communicate it only through validated HMI output; it cannot override, negotiate, or delay the action.

6. Runtime Governance and Degradation-Aware Operation

6.1. Motivation

An AI system intended for safety-relevant contexts cannot be considered sufficiently trustworthy on the basis of average-case performance alone; runtime governance is also required. The transition-of-control context is particularly demanding: the system must remain acceptable precisely when conditions are most challenging, i.e., under sensor degradation, data staleness, adversarial inputs, and time pressure. In this setting, degradation does not affect only the vehicle's control capability; it also affects the validity of the assistance policy presented to the driver. A companion AI component that reasons over stale, conflicting, or poorly grounded state estimates may select the wrong timing, the wrong modality, or the wrong communicative register even if the underlying control stack remains operational. This section therefore focuses on the governance mechanisms required to keep adaptive transition support bounded, inspectable, and safe under degraded conditions.

6.2. Runtime Watchdog and Policy Enforcement

The architecture includes a dedicated runtime watchdog that continuously validates the companion AI component's proposed outputs against a set of formalized TCAI policies [85,86]. These policies define the permissible action space of the companion AI component, specify constraints on interaction intensity as a function of driver state, enforce freshness thresholds on state estimates, and define the conditions under which the companion AI component must defer to deterministic fallback procedures.

The watchdog operates at a faster timescale than the companion AI component's reasoning loop. It receives the companion AI component's proposed interaction actions *before* they are delivered by the HMI and evaluates whether each proposed action remains within the policy boundary. Actions that violate policy constraints, for example, a high-intensity haptic intervention proposed when the driver is already in an acute stress state or an explanatory interaction that is too verbose for the current urgency level, are blocked and replaced with a safe default. In this sense, the watchdog does not merely constrain content; it also constrains phase transitions, escalation timing, and interaction style.

Policy violations are logged with the full state context that triggered them, supporting post hoc analysis and policy refinement. The watchdog itself uses only deterministic formally specified logic; it contains no learned components. This design follows the broader principle in runtime-assurance research that the safety supervisor must be formally verifiable, even when the supervised component is not [85].

6.3. Veto Handling Levels and State-Machine Logic

The watchdog should implement graded veto handling rather than binary accept/reject logic [85,86]. Each companion AI proposal is evaluated against the semantic state interface snapshot, admissible-action policy, urgency, uncertainty and consistency flags, and advisory/control boundary. The resulting veto state can be conceptualized as a five-level state-machine-style ladder, illustrated in Figure 3:

- **V0— Accept.** The output is policy-compliant, grounded in fresh and sufficiently confident semantic state, appropriate for the current urgency, driver state, and interaction phase, delivered through the adaptive HMI, and logged.
- **V1—Sanitize.** The interaction is admissible, but wording, framing, or uncertainty expression is not fully compliant. The watchdog removes unsupported claims, adds

uncertainty qualifiers, suppresses anthropomorphic or over-authoritative phrasing, or replaces free-form language with a guarded template.

- **V2—Downgrade.** The objective is admissible, but modality, salience, intensity, or explanation depth is inappropriate. The watchdog may reduce explanation depth, replace speech with a shorter cue, lower haptic intensity, or defer elaboration to the post-transition phase.
- **V3—Block.** The output violates policy, lacks semantic support, attempts to negotiate a safety boundary, gives false reassurance, or conflicts with state-consistency flags. It is replaced with a pre-verified safe default or guarded transition-support message.
- **V4—Deterministic fallback.** If no valid advisory output can be produced, the companion AI component is unavailable, repeated proposals are rejected, or the state snapshot is too stale, inconsistent, or degraded for personalized assistance, the TCAI layer invokes deterministic fallback. This fallback remains independent of the companion AI component and cannot modify vehicle control or MRC execution.

The transition logic is conservative: the watchdog checks state freshness and consistency, action admissibility, modality/intensity constraints, then explanation depth and wording. A proposal can move directly to a more restrictive veto state when urgency, uncertainty, or violation severity requires it. Recovery requires a newer, more consistent semantic state interface snapshot and a proposal satisfying the relevant checks. All veto transitions are logged with the rejected proposal, active semantic state, violated policy, and selected replacement or fallback action.

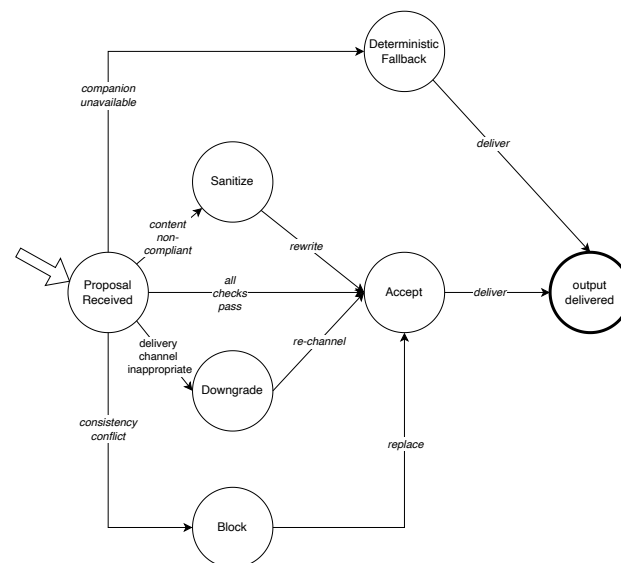


Figure 3. Graded veto-handling state-machine logic of the runtime watchdog.

6.4. Degradation-Aware Reasoning

Building on the state-consistency mechanisms introduced in Section 5.6, the TCAI layer is designed to reason explicitly about the quality of its own inputs. Different degradation patterns require different behavioral responses because they do not all compromise the assistance policy in the same way.

Missing data. If a state variable is absent from the semantic state interface, the TCAI defaults to a conservative estimate for that variable. A missing HDT readiness estimate, for instance, is treated as indicating low-to-moderate readiness rather than readiness. The design principle is that absence of evidence should not be interpreted as evidence of adequacy.

Stale estimates. State estimates that exceed their configured freshness thresholds trigger a reduction in the companion AI component's confidence in that dimension under TCAI governance and a corresponding increase in the conservatism of the assistance policy. The severity of this adjustment is policy-configurable and depends on both the degree of staleness and the current urgency level.

Asynchronous or conflicting estimates. When state estimates from different sensing modalities or digital-twin sub-models are asynchronous or mutually inconsistent, the TCAI applies a conservative arbitration rule: pending resolution, the more cautious estimate is used. This does not imply that the selected estimate is always correct; rather, it reflects the principle that transition support should fail toward bounded caution when key semantic inputs disagree. Such cases are logged for post hoc analysis and model refinement.

Adversarial or anomalous inputs. The architecture includes an anomaly-detection layer at the semantic state interface that flags inputs deviating from expected operational ranges or exhibiting patterns inconsistent with valid sensor behavior. Flagged inputs are either rejected in favor of the most recent valid estimate or trigger a degraded-mode transition protocol depending on policy and urgency.

Taken together, these mechanisms ensure that degradation is handled not only as a perception or estimation problem but as a condition that directly modulates how much confidence the system should place in its own transition assistance policy.

6.5. Deterministic Fallback Separation

If the companion AI component becomes unavailable, is suspended by the TCAI layer's watchdog, or produces outputs that cannot be validated, the system falls back to a deterministic rule-based transition protocol. This protocol is independent of the companion AI component and is invoked by the TCAI layer's runtime governance; it implements the minimum required transition-support behavior without any learned or adaptive logic [86,87]. The deterministic fallback does not attempt to replicate the companion AI component's graduated or personalized assistance; it delivers a standardized policy-compliant transition sequence designed to remain valid even at the boundary of the companion's operational envelope.

This separation ensures that the system has well-defined safe behavior when the companion AI component can no longer be trusted to support the transition appropriately. In other words, fallback safety does not depend on continued confidence in adaptive reasoning.

6.6. Uncertainty Communication

When the companion AI component's confidence in a transition-support proposal is low because of degraded inputs, it should communicate this uncertainty to the driver in a driver-appropriate way rather than conceal it. This principle extends trustworthiness-by-design to the interaction layer: a companion AI component that presents uncertain assessments with false confidence undermines calibrated trust more severely than one that communicates appropriate hedging [54,84]. In transition contexts, uncertainty communication is therefore not merely a matter of interface style; it is part of trust calibration itself.

The practical design of uncertainty communication, including what information to surface, at what level of detail, and in what modality, is still an open research question. In particular, future work must determine when uncertainty should be made explicit, when it should remain implicit in a more conservative interaction policy, and when explanation should be reduced or suppressed altogether in favor of brief directive guidance. For this reason, uncertainty communication should be treated as a priority validation target of the proposed architecture.

7. Research Agenda and Validation Dimensions

This paper presents a motivated conceptual architecture and research roadmap; it does not present completed experiments, user studies, or implementation evaluations. We explicitly frame the following as an agenda for future research. More specifically, the proposed program requires a combination of simulator studies, cyber-range-like perturbation campaigns, and longitudinal human factor evaluation because transition quality in this setting depends not only on driving performance but also on trust calibration, stress containment, explanation timing, and robustness under degraded conditions.

A first validation protocol would compare a conventional TOR-only or fixed-alert baseline with the proposed digital-twin-mediated TCAI-governed companion AI support in a repeated-measures driving-simulator study [2,8]. Scenario factors should include planned ODD exits, urgent handovers, and degraded or conflicting sensing; driver-state factors should include rested/attentive and passive-fatigue or disengagement conditions [11,12]. Dependent variables should combine takeover-quality measures, such as takeover time, SDLP, steering variability, braking smoothness, TTC, jerk, gaze-road recovery, and post-handover stabilization time, with workload, stress/arousal, explanation comprehension, trust calibration, and perceived appropriateness [7,9,33,35,38]. This design would test the fusion model, phase-selection policy, and watchdog-governed interaction strategy against a conventional baseline.

7.1. Dimension 1: Planned Transition Quality

The foundational validation dimension addresses whether the proposed graduated human-aware transition process improves handover quality relative to conventional TOR approaches under planned transition conditions.

This validation dimension focuses specifically on the quality of planned transitions of control under graduated human-aware support conditions. Within this perspective, *transition quality* should not be reduced to takeover latency alone since extremely short takeover times may also reflect premature motor reactions performed without sufficient situational reorientation. Instead, transition quality should be evaluated as a function of prior preparation level, urgency, and driver state through a combination of dynamic driving performance, safety-related indicators, perceptual–motor re-engagement measures, and user-centered experience metrics.

Metrics of interest include:

- **Dynamic vehicle-control stability**, including standard deviation of lane position (SDLP), steering-wheel variability, lateral acceleration, braking smoothness, post-takeover stabilization time, and other indicators of whether manual control is resumed in a controlled and coordinated manner.
- **Surrogate safety measures** (SSMs), including minimum time-to-collision (TTC), collision or near-miss occurrence, and resultant jerk, to assess the objective safety margins maintained during the immediate post-transition interval.
- **Perceptual–motor re-engagement measures**, including gaze redistribution toward the road scene, gaze–steering coordination, mirror-check behavior, and the temporal relationship between guiding fixations and steering actions, as indicators of progressive re-coupling between perception and manual control.
- **Situational awareness and cognitive readiness indicators**, including SAGAT-like probes where experimentally compatible, behavioral evidence of environmental monitoring, and measures related to the driver’s understanding of the ongoing transition context.
- **Direct and indirect user-experience measures**, including subjective workload scales, perceived transition smoothness, perceived appropriateness of interaction timing,

physiological indicators of stress or cognitive activation (e.g., HRV and GSR), and overall acceptance of the transition-support process.

7.2. Dimension 2: Time-Constrained and Urgent Transitions

A critical test of any transition-support architecture is its behavior under time pressure. The companion AI component, under TCAI governance, must not attempt elaborate preparation sequences when a transition must occur in two seconds. The governance architecture must enforce appropriate response-mode selection under urgency, including the ability to suppress explanation when explanatory interaction would consume time or working-memory capacity needed for immediate control resumption. Validation should examine whether the system correctly escalates to its deterministic fallback protocol in time-constrained conditions, whether the interaction intensity under urgency is appropriate rather than counterproductive [88], and whether the boundary between anticipatory support and directive execution is selected at the right moment.

7.3. Dimension 3: Degraded and Adversarial Conditions

This dimension tests the robustness of the degradation-aware reasoning described in Section 6. Scenarios of interest include: camera-based DMS signal loss during a transition event, conflicting estimates from gaze and physiological channels, stale HD-map data at a previously unknown ODD boundary, and adversarial perturbation of environmental sensing inputs. These conditions should not be treated as marginal edge cases but as a distinctive validation dimension of the proposed architecture.

Validation metrics should address whether the system correctly activates conservative fallback behavior, whether it appropriately communicates uncertainty to the driver, whether false-safe (unnecessarily conservative) behavior remains bounded, and whether degraded conditions lead to safe adaptation of the assistance policy rather than silent overconfidence. Cyber-range-style fault injection and perturbation campaigns are especially relevant here because they allow systematic testing of watchdog logic, interface staleness handling, and uncertainty-aware escalation under controlled conditions. Digital-twin-based simulation, sim2real, and parallel-intelligence approaches for autonomous driving provide a complementary methodological basis for testing transition-support policies under controlled but realistic degraded conditions [89].

7.4. Dimension 4: Soft Vigilance and Stress Regulation

The motivating scenario in Section 4.1 suggests a distinct validation dimension: whether graduated pre-transition interaction aimed at inducing soft vigilance reduces driver stress at transition onset and improves takeover quality relative to unmediated alerts. In this framework, soft vigilance is operationalized as a state of “relaxed but alert” cognitive reorientation, acting as a functional bridge between deep disengagement (passive fatigue) and full operational readiness.

Investigating these constructs will require the exploration of multimodal operationalization strategies capable of rendering both soft vigilance and transition-related stress measurable under different driving and interaction conditions. One promising avenue for operationalizing soft vigilance [12] involves the use of eye-tracking indicators such as gaze entropy, fixation-duration patterns, and visual behavior over critical areas of interest (AOIs), with the aim of detecting transitions from passive environmental exploration to more active situational scanning and road-oriented monitoring activity. Stress may be explored through physiological indicators associated with sympathetic activation and cognitive load. Following Markelius et al. [35], relevant measures may include Baevsky’s stress index (SI), heart-rate variability (HRV) parameters such as RMSSD and SDNN, gal-

vanic skin response (GSR/EDA), and related autonomic indicators capable of capturing transition-related arousal dynamics in real time.

The relationship between these physiological and behavioral indicators and the resulting quality of takeover should be investigated explicitly considering the influence of contextual factors such as road scenarios and the characteristics of the HMI design itself, including interaction timing, communication modality, trust-calibration strategies, and contextual explainability mechanisms.

Future work should also establish the boundary between beneficial soft-vigilance induction and counterproductive over-involvement: a driver who is made too alert too early may experience unnecessarily elevated arousal before the transition is actually required. In the same vein, stress regulation should be evaluated not only as stress reduction but as *stress containment at an appropriate level for the upcoming transition*.

7.5. Dimension 5: Trust Calibration Trajectory

Future research should investigate how driver trust and behavioral adaptation evolve through repeated control transitions and across varied operational conditions. This longitudinal perspective is essential to determine whether interaction strategies support appropriately calibrated reliance over time, specifically when compared to conventional TOR-only architectures that lack explicit calibration mechanisms [33,38]. Within this framework, trust must be treated not as a static post hoc score but as a dynamic multi-layered process shaped by accumulated experience and system transparency [18,29]. Particular attention should be devoted to exploring how drivers adjust their mental models after the system communicates uncertainty, reduced capability, or the necessity for a fallback [90]. Researchers should examine the “anchoring effect” of initial interactions as early exposure to (in)accurate system information can create persistent cognitive biases that influence future takeover decisions and trust stability [57]. A well-calibrated interaction strategy should facilitate an appropriate adjustment of driver expectations without inducing either chronic over-reliance, which leads to complacency and reduced monitoring [91], or a complete loss of confidence following system degradation events [30,57]. Future validation should also test whether the Trust Dynamics Module distinguishes calibrated reliance, over-trust, and under-trust trajectories and whether explanation-style adaptations improve calibration without becoming persuasive or manipulative.

7.6. Dimension 6: Explainability Effectiveness

A critical validation dimension concerns whether contextualized transition-related explanations improve driver understanding, situational reorientation, and takeover quality compared to generic notifications [8,34]. Explainability effectiveness should not be evaluated solely through subjective satisfaction or perceived transparency; it must be operationalized through measurable outcomes, including mental-model accuracy and completeness, understanding of system limitations, and SA recovery [63,90]. Following the situation awareness-based agent transparency (SAT) framework [62,63], we propose evaluating how different levels of transparency, ranging from basic status updates (Level 1) to detailed underlying logic (Level 2) and future end-state projections (Level 3), influence the driver’s ability to re-engage with the driving task. However, assessments must also account for the “transparency paradox”, where high-specificity explanations may inadvertently increase driver anxiety if they reveal inconsequential perception errors [32]. Particular attention should be devoted to how the utility of explanatory content evolves across transition phases. Although richer causal explanations can support trust calibration and mental-model development during anticipatory stages, evidence suggests that “more explicit is not always better” during urgent takeovers [23]. In high-complexity traffic

scenarios, brief advisory messages are often more effective than detailed action guidance, which may impose excessive cognitive workload or induce interference [23]. Future studies should therefore use multimodal evaluation strategies integrating objective SA measures with physiological stress indicators to investigate how explanation timing, level of detail, and contextual framing influence transition performance under realistic temporal and cognitive constraints [25,35]. Future validation should test whether the E0–E3 explanation-depth policy selects appropriate detail across urgency, cognitive load, task complexity, and mental-model mismatch.

7.7. Open Research Questions

The following are specific first-order open questions for the proposed research program:

- What is the minimum state-variable set and freshness threshold required for adequate companion AI decision quality under TCAI governance across the identified transition scenario classes?
- How should soft vigilance be operationally defined, measured, and thresholded for use as an HDT state variable in real-time takeover-support decisions?
- What pre-transition interaction lead times, interaction styles, and contextual anchors most effectively induce soft vigilance from a passive fatigue baseline without imposing counterproductive cognitive load [12]?
- How should the three-phase communication strategy be timed and calibrated for drivers with heterogeneous stress baselines and automation familiarity levels?
- What formal specification and verification methods are appropriate for the TCAI layer's action-space policies and watchdog logic?
- What are the minimal requirements for a deterministic fallback protocol to remain acceptably safe across the scenario distribution?
- At what point does XAI-mediated trust in the anticipatory phase become counterproductive under very short time-to-ODD-exit conditions, and how should the system detect and respond to that boundary [34]?
- How should driver veto, partial disabling, and conservative policy fallback be validated so that user control does not undermine safety-critical transition support?

8. Technical Development Roadmap

The validation agenda in Section 7 addresses empirical evaluation; this section presents a conceptual roadmap rather than an implementation specification. It distinguishes trainable state-estimation components, formally specified policy and phase-selection components, the assumed LLM-based companion AI component, and deterministic watchdog and fallback components that should remain non-learned.

8.1. Technical Building Blocks and Semantic Outputs

HDT-centered estimation may rely on multimodal evidence from gaze behavior, PERCLOS, blink rate, fixation duration, gaze dispersion, HRV/RMSSD, EDA/GSR, pupil dilation, facial expression, SDLP, steering variability, and interaction history. These inputs can support probabilistic or temporal estimators, including Bayesian models, DBNs, HMMs, LSTM or temporal-transformer classifiers, and multimodal fusion approaches [69,92–94]. The intended outputs are semantic variables such as vigilance level, fatigue index, soft-vigilance estimate, cognitive-load level, stress indicator, readiness estimate, trust profile, and latent mental-model estimate, each accompanied by uncertainty where possible.

A candidate HDT fusion scheme can be formulated as temporal probabilistic evidence fusion [69,75,76]. At each time step, a latent driver-state vector, including readiness, fatigue, cognitive load, soft vigilance, stress/arousal, and trust-related state, is updated from

multimodal observations weighted by predictive relevance and by confidence, freshness, and provenance metadata exposed through the semantic state interface [92–94]. DBNs, HMMs, Bayesian filters, or calibrated temporal neural estimators may be used to preserve uncertainty; missing, stale, or conflicting observations increase state uncertainty and may trigger conservative arbitration or degraded-mode policies. This fusion family is an illustrative operationalization, not an exclusive modeling choice, but it fits the architecture because the relevant driver states are latent, time-varying, partially observable, and affected by heterogeneous sensing quality.

Vehicle and context digital twins would complement HDT estimates by representing automation status, ODD-boundary proximity, sensor health, perception confidence, road geometry, traffic, weather, and transition urgency. The semantic state interface should expose these outputs with confidence, freshness, and provenance metadata so that the TCAI layer reasons over bounded auditable state descriptions rather than unconstrained sensor data.

Phase selection and admissible-action policy should be formally specified through finite-state machines, decision tables, or symbolic policies rather than learned implicitly. TCAI-mediated companion AI output should rely on constrained prompting, guarded templates, verified knowledge resources, and output validation, while runtime governance should rely on a deterministic watchdog that validates, blocks, or redirects proposed outputs before HMI delivery. EEG/BCI signals may be investigated as optional future supplementary inputs, but they are not required by the current architecture because of stability, artifact, privacy, and acceptability concerns.

8.2. Driver Mental-Model Representation

The driver's mental model is latent and only partially observable, so the companion AI component should not claim direct access to it. Instead, the HDT should maintain a probabilistic estimate of model calibration from observable proxies.

Candidate proxies include compliance or non-compliance with prompts and handover requests, clarification requests, takeover quality and latency, response to contextualized explanations, gaze patterns toward the system-status display, and longitudinal trust trajectory.

Together, these proxies form an observational record from which a latent mental-model accuracy index can be estimated using a Bayesian longitudinal model or recurrent latent-variable estimator. This estimate is uncertain and should reach the companion AI component through the TCAI layer with an explicit confidence score. Subject to TCAI constraints, persistent miscalibration may justify more explicit corrective explanations during anticipatory preparation, whereas good calibration may justify concise contextual updates.

8.3. Training, Calibration, and Formal Specification Strategy

Not all framework components require the same development strategy. Driver-state estimators for fatigue, cognitive load, vigilance, stress, and readiness require training or calibration from labeled simulator, naturalistic-driving, physiological, and behavioral data, with attention to ground-truth quality, cross-session variation, and population differences. Trust and mental-model estimators likewise require longitudinal calibration from interaction logs, clarification requests, takeover behavior, explanation responses, and trust-calibration probes because the constructs are latent and indirect.

By contrast, adaptive phase-selection policies should be formally specified rather than learned. Their role is to select the active transition phase and admissible action set under governance requirements, so finite-state machines, decision tables, symbolic policies, policy-correctness audits, and scenario-coverage tests are more appropriate than opaque learned policies.

The companion AI component is assumed to be LLM-based, but this paper does not train, benchmark, fine-tune, or compare such models. The contribution concerns the TCAI layer that constrains and governs that component through bounded reasoning, policy enforcement, constrained prompts, guarded templates, verified knowledge resources, and output validation against the current semantic state interface snapshot. Watchdog and fallback mechanisms should remain deterministic, non-learned, independently specified, and separated from both the companion AI component and vehicle-control functions.

8.4. Staged Development Roadmap

A plausible development path would proceed in stages, each adding integration complexity only after the relevant conceptual and governance requirements are explicit. The sequence is indicative, not a claim of completed work.

1. **S1: Component development.** Develop candidate HDT state estimators for fatigue, cognitive load, vigilance, readiness, trust, and mental-model calibration while separately specifying phase-selection policy and watchdog logic.
2. **S2: Integration and semantic state interface.** Integrate HDT, VDT, and CDT outputs through the semantic state interface, including confidence, freshness, provenance, and degradation metadata for bounded downstream reasoning.
3. **S3: TCAI-mediated companion AI integration.** Place the assumed LLM-based companion AI component behind TCAI guardrails, constrained prompting, verified explanation resources, and deterministic watchdog validation; evaluate explanation comprehension and trust calibration in simulator studies.
4. **S4: Phase-sensitive HMI integration.** Connect the TCAI layer to an adaptive HMI and evaluate the three-phase interaction strategy across planned, anticipatory, and urgent transition scenarios.
5. **S5: Longitudinal and robustness validation.** Examine trust-calibration trajectories, soft-vigilance support, stress regulation, sensing degradation, and adversarial or out-of-distribution inputs over repeated simulator sessions.
6. **S6: Instrumented-vehicle study.** If simulator evidence supports further development, assess the framework in an instrumented vehicle under closed-track or regulated public-road conditions to study ecological validity and generate safety-case inputs.

8.5. A Note on Over-Claiming

The roadmap describes what would be required to develop and validate a full implementation of the proposed digital-twin-mediated framework. This paper does not claim that such an implementation exists, that the architecture has been empirically validated, or that a new companion AI model has been designed, trained, benchmarked, or compared. The roadmap shows technical coherence and a credible path to realization, not a completed path.

9. Ethical and Interaction Design Considerations

The deployment of a conversational adaptive companion AI component governed by a TCAI layer in safety-relevant driving transitions raises ethical challenges that are not separable from architecture and interaction design. In the present vision, the central ethical requirement is not to make the system appear helpful at all costs but to ensure that its helpfulness remains bounded, non-deceptive, and compatible with calibrated trust.

9.1. Over-Reliance and Automation Complacency

A major ethical risk of the proposed TCAI-governed companion AI component is the very risk it is intended to mitigate: over-reliance. A sufficiently capable and consistently

helpful companion AI component may, over time, encourage drivers to become less self-prepared for transitions precisely because they expect the system to manage their readiness. This risk is not theoretical. Research on automation complacency shows that reliable assistance can reduce vigilance and increase dependence [95,96]. The TCAI layer must therefore govern the companion AI component not only to support transitions but also to avoid deepening the driver's structural disengagement from the driving context.

This creates a fundamental design tension. A companion AI component that is useful enough to improve transition quality may also become convenient enough to erode the driver's own preparation if poorly designed. For this reason, the objective of the companion AI component under the TCAI layer should not be to maximize comfort or passive dependence but to support readiness while preserving the driver's role as the residual human agent in the transition loop.

9.2. Trust Manipulation, Anthropomorphic Framing, and Negotiation

Conversational and natural-language interaction in a safety-relevant context raises the risk of anthropomorphic projection: drivers may attribute human-like reliability, understanding, or judgment to a system that remains bounded, uncertain, and fallible. Research on human-robot and human-vehicle interaction suggests that natural-language interfaces can amplify this effect [97,98]. An interaction design that implicitly exploits anthropomorphic tendencies to increase compliance with transition prompts is therefore not merely stylistically questionable; it can produce miscalibrated trust that may offset, or even outweigh, the intended safety benefits of the companion AI component.

A related risk is that users may attempt to negotiate with the system as if it were a flexible conversational partner rather than a bounded transition support mechanism. Research on passenger-vehicle LLM interaction has already documented cases in which users attempt to persuade the system to violate policy constraints or simulate exceptional circumstances to obtain prohibited behaviors [99]. In the present setting, this risk is especially salient during anticipatory phases, where the interaction is necessarily more contextualized and verbal. The response, however, should not be to eliminate conversational support entirely but to ensure that runtime governance, watchdog enforcement, and interaction framing do not invite negotiation as a mode of control.

Accordingly, the companion AI component should, as required by the TCAI layer's interaction policies, consistently communicate its bounded and non-human status, avoid language implying human-level situational awareness, and express system limitations when relevant. This is particularly important for soft-vigilance support: personalized prompts may be useful for passive fatigue management [12], but personalization must not drift into pseudo-intimacy or rhetorical reassurance that inflates perceived system competence beyond what is warranted.

9.3. False Reassurance and the Duty of Uncertainty Communication

Closely related is the problem of false reassurance. A companion AI component that presents a situation as manageable when its own confidence is low risks impairing driver decision-making. We therefore argue that the system has a design-level duty to communicate uncertainty honestly and to behave conservatively when confidence is insufficient. When the HDT readiness estimate is unavailable or low-confidence, the appropriate response is not to proceed as if the driver is ready; it is to communicate the limitation and adjust the transition policy accordingly.

This principle creates a genuine design tension. Interactions should remain clear, lightweight, and non-disruptive, yet they must not conceal the fact that the system is operating under degraded knowledge. Research on uncertainty communication suggests

that naive presentation of numerical confidence values is often counterproductive and may itself increase cognitive load [84]. The challenge is therefore not simply to “show uncertainty” but to do so in a way that preserves calibrated trust without undermining the driver’s ability to act.

9.4. Privacy, Data, and Driver Profiling

The HDT accumulates longitudinal information about the driver’s physiological state, trust history, and behavioral patterns. This creates a significant data-collection footprint and raises clear privacy concerns. The minimum viable set of personal data required for effective transition support should be used, retention periods should be explicitly bounded, and drivers should retain meaningful control over profiling-related data. These considerations reinforce the architectural preference for edge-first processing and privacy-preserving inference introduced earlier. Driver-facing control over optional profiling and personalization is discussed further below.

The regulatory implications are also non-trivial. The use of driver-related physiological, behavioral, and profiling data for safety purposes must be assessed against the applicable legal framework for high-risk AI systems and data protection in automotive contexts. In particular, future deployment would need to account not only for requirements related to data governance, transparency, and human oversight but also for post-market monitoring obligations under the EU AI Act [87].

9.5. Interaction Design Must Not Conceal Limitations

The overarching ethical principle for the companion AI component’s interaction design is that natural interaction must not conceal system limitations. Outputs should not be framed as authoritative unhedged assessments when they are in fact based on incomplete, uncertain, or degraded state information. The interaction design must therefore communicate, with appropriate restraint, what the system knows, what it is uncertain about, and what it cannot know. This is not merely an ethical nicety; it is a functional requirement for maintaining calibrated trust across repeated driver–system interaction [53,54].

From this perspective, boundedness is not only an architectural property but also an interaction-design obligation. A trustworthy companion AI is not one that sounds the most confident or the most human-like but one that remains consistently aligned with its real epistemic limits and policy constraints.

9.6. Driver Veto and Partial Disabling Mechanisms

A trustworthy interaction design should define what the driver may veto or disable and what must remain active for safety. Because the companion AI component is advisory, the driver should be able to partially disable non-essential functions, such as conversational style, personalized phrasing, extended explanations, non-critical soft-vigilance prompts, and optional profiling features not required for immediate safety. These controls should be available through clear HMI controls, not negotiation with the companion AI component.

Partial disabling should be conditional: it should be allowed only when the vehicle is not in imminent handover, high-urgency degradation, or active fallback and when semantic state indicates reduced support does not create immediate transition risk. Disabling personalization or extended explanation should trigger a conservative less personalized policy: shorter validated prompts, guarded templates, reduced conversational flexibility, and earlier fallback invocation under uncertainty.

Several functions remain non-disableable: mandatory TOR signaling, safety-critical warnings, MRC execution, deterministic fallback, runtime watchdog validation, advisory/control boundary enforcement, and minimum state monitoring required to decide

whether a companion AI output can be safely delivered. These are part of the safety and governance envelope, not discretionary conversational assistance.

Driver veto events should themselves be logged as part of the interaction history, subject to the data-minimization and retention constraints described above. This helps the HDT and Trust Dynamics Module to distinguish rejection of a specific interaction style from general non-compliance, overload, or under-trust. A veto should not license negotiation, persuasion, or repeated requests for disabled functions; the system should acknowledge the preference, apply the conservative policy, and re-enable richer assistance only through explicit driver action or predefined safety-relevant conditions.

9.7. Scope and Limitations

The present work is a conceptual architecture and technical development roadmap, not an implementation or validation study. The proposed digital-twin-mediated framework should be read as a structured research hypothesis with a coherent but unvalidated technical proposal. No prototype, empirical evaluation, or component-level benchmark is reported here; this scope boundary is intended to support the experimental program needed to validate the framework.

The literature discussion is selective and purpose-driven rather than systematic. It motivates the integration gap and proposed architectural response and situates the contribution relative to HASs, classical cognitive assistance, and emerging LLM-in-vehicle research. Readers seeking systematic coverage should consult dedicated reviews [2,25,28].

Several elements remain open research problems before implementation: real-time measurement of soft vigilance as an HDT variable; calibration of driver mental-model estimates from interaction logs; explanation timing and depth under varying urgency and cognitive load; formal specification of runtime-governance policies; and sim-to-real transfer of driver-state estimators.

Section 8 identifies these problems and proposes a staged development program. EEG and BCI signals are included for conceptual completeness but are treated as possible future signals rather than required components of the current architecture.

Finally, the proposed architecture assumes that appropriate in-cabin sensing (camera-based gaze tracking and, where available, physiological sensing) is present. The performance of the HDT, and therefore of the companion AI component's reasoning, depends on the quality and availability of these inputs. Sensing unavailability is handled through the degradation-aware reasoning mechanisms described in Section 6, but a companion AI component operating with severely degraded sensing will necessarily provide less personalized and less precisely calibrated assistance. This dependency should be made explicit in any future safety case or deployment specification.

10. Conclusions

Takeover support in automated driving remains a demanding human–automation interaction problem, but conversational, generative, or agentic-style AI assistants change its architectural character. The issue is no longer only how early or loudly to issue a takeover request (TOR) but how to support the driver through bounded state-grounded phase-sensitive interaction while preserving a hard separation between advisory assistance and safety-critical vehicle control.

This paper has proposed a digital-twin-mediated framework for that problem. Foundational work on human assistance systems [67], cognitive assistance [68], and driver-fatigue recognition using dynamic Bayesian networks [69] already established the basis for adaptive state-aware human support. The present contribution is architectural: it asks how an

assumed LLM-based companion AI component can support takeover interaction without becoming ungrounded, unbounded, or safety-critical.

The framework grounds advisory interaction in human, vehicle, and context/road digital twins, mediated through a semantic state interface carrying confidence, freshness, and provenance metadata. The TCAI layer is the architectural trust layer that grounds, constrains, supervises, and governs the companion AI component; it is not a newly trained model and not the whole system. The companion AI component may select prompts, explanations, and HMI actions, but it does not actuate the vehicle, has no write-access to vehicle-control parameters, and remains separated from the vehicle-control and minimum risk condition stack.

The proposed hybrid technical realization concerns the digital-twin-mediated framework and TCAI layer, not alternative architectures for the assumed LLM-based companion AI component. Probabilistic temporal models estimate driver state; symbolic or rule-based policies perform phase selection and admissible-action control; TCAI-mediated controlled generation supports explanations; and a deterministic watchdog validates outputs and triggers fallback. This keeps phase selection, policy enforcement, watchdog supervision, fallback, and driver mental-model estimation outside open-ended generative components.

Three interaction ideas remain central: soft vigilance as an intermediate state between passive disengagement and readiness; *XAI-mediated trust calibration* to align the driver's inferred mental model with system condition; and *phase-sensitive communication* that moves from low-intrusion contextual support to anticipatory explanation to brief directive guidance as urgency increases.

Although motivated primarily by SAE Level 3 automation, the framework may also inform degraded or fallback-related Level 4 scenarios where human and automated agency must be managed under uncertainty. The roadmap identifies development and validation directions for driver-state estimation, mental-model inference, explanation generation, phase-sensitive HMI integration, runtime governance, and trust-calibration evaluation. The value of the framework is not to replace vehicle control but to make takeover support more cognitively appropriate, contextually intelligible, and architecturally bounded as in-vehicle interaction evolves toward general-purpose AI assistance.

Author Contributions: Conceptualization, R.P. and R.G.; methodology, R.P. and R.G.; investigation, R.P., F.D.S., L.B. and R.G.; writing—original draft preparation, R.P. and R.G.; writing—review and editing, R.P., F.D.S., L.B. and R.G.; visualization, R.G.; supervision, R.P. and R.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: FigureLabs.ai (<https://figurelabs.ai/>, accessed on 26 May 2026) was used to support the preparation of limited graphical elements in Figure 1. The conceptual design, scientific interpretation, and final validation of the figure, as well as the manuscript content, were carried out by the authors. The authors have reviewed and approved the final content and take full responsibility for the publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. SAE International. *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*; Technical Report J3016_202104; SAE International: Warrendale, PA, USA, 2021. https://doi.org/10.4271/j3016_201609.
2. Sekadakis, M.; Yannis, G. Systematic review and meta-analysis of take-over time from automated driving at SAE levels 2 and 3 to manual control. *Transp. Res. Part F Traffic Psychol. Behav.* **2025**, *113*, 263–306. <https://doi.org/10.1016/j.trf.2025.04.003>.
3. Gerber, M.A.; Schroeter, R.; Ho, B. A human factors perspective on how to keep SAE Level 3 conditional automated driving safe. *Transp. Res. Interdiscip. Perspect.* **2023**, *22*, 100959. <https://doi.org/10.1016/j.trip.2023.100959>.
4. Guo, H.; Qiu, H.; Zhou, Y.; Deng, Y. Analysis of Influencing Factors of Level 3 Automated Vehicle Takeover: A Literature Review. *Sustainability* **2024**, *16*, 8345. <https://doi.org/10.3390/su16198345>.
5. Endsley, M.R.; Kiris, E.O. The out-of-the-loop performance problem and level of control in automation. *Hum. Factors* **1995**, *37*, 381–394. <https://doi.org/10.1518/001872095779064555>.
6. Merat, N.; Seppelt, B.; Louw, T.; Engström, J.; Lee, J.D.; Johansson, E.; Green, C.A.; Katazaki, S.; Monk, C.; Itoh, M.; et al. The “Out-of-the-Loop” concept in automated driving: Proposed definition, measures and implications. *Cogn. Technol. Work* **2018**, *21*, 87–98. <https://doi.org/10.1007/s10111-018-0525-8>.
7. Gold, C.; Körber, M.; Lechner, D.; Bengler, K. Taking Over Control From Highly Automated Vehicles in Complex Traffic Situations: The Role of Traffic Density. *Hum. Factors J. Hum. Factors Ergon. Soc.* **2016**, *58*, 642–652. <https://doi.org/10.1177/0018720816634226>.
8. Kim, S.; Egmond, R.; Happee, R. Effects of user interfaces on take-over performance: A review of the empirical evidence. *Inf.* **2021**, *12*, 162. <https://doi.org/10.3390/info12040162>.
9. Zeeb, K.; Buchner, A.; Schrauf, M. What determines the take-over time? An integrated model approach of driver take-over after automated driving. *Accid. Anal. Prev.* **2015**, *78*, 212–221. <https://doi.org/10.1016/j.aap.2015.02.023>.
10. Feldhütter, A.; Gold, C.; Schneider, S.; Bengler, K. How the duration of automated driving influences take-over performance and gaze behavior. In *Proceedings of the Advances in Ergonomic Design of Systems, Products and Processes*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 309–318. https://doi.org/10.1007/978-3-662-53305-5_22.
11. Cabrall, C.; Happee, R.; Winter, J.D. From Mackworth’s clock to the open road: A literature review on driver vigilance task operationalization. *Transp. Res. Part F Traffic Psychol. Behav.* **2016**, *40*, 169–189. <https://doi.org/10.1016/J.TRF.2016.04.001>.
12. Yu, Y.; Cockram, L.; Pardo, J.; Li, X.; Rakotonirainy, A.; Kuo, J.; Demmel, S.; Lenné, M.; Schroeter, R. Leveraging LLM-based Conversational Agents to Combat Passive Fatigue in Conditional Automated Driving Contexts. In *Proceedings of the International Conference on Automotive User Interfaces and Interactive Vehicular Applications*; ACM: New York, NY, USA, 2025; pp. 126–130. <https://doi.org/10.1145/3744335.3758493>.
13. Bouhsissin, S.; Sael, N.; Benabbou, F. Driver behavior classification: A systematic literature review. *IEEE Access* **2023**, *11*, 14128–14153. <https://doi.org/10.1109/ACCESS.2023.3243865>.
14. Merat, N.; Jamson, A.H.; Lai, F.C.H.; Daly, M.; Carsten, O.M.J. Transition to manual: Driver behaviour when resuming control from a highly automated vehicle. *Transp. Res. Part F Traffic Psychol. Behav.* **2014**, *27*, 274–282. <https://doi.org/10.1016/j.trf.2014.09.005>.
15. Casner, S.M.; Hutchins, E.L.; Norman, D. The Challenges of Partially Automated Driving. *Commun. ACM* **2016**, *59*, 70–77. <https://doi.org/10.1145/2830565>.
16. Bainbridge, L. Ironies of automation. In *at - Automatisierungstechnik*; Elsevier: Amsterdam, The Netherlands, 1982; pp. 129–135. [https://doi.org/10.1016/0005-1098\(83\)90046-8](https://doi.org/10.1016/0005-1098(83)90046-8).
17. Parasuraman, R.; Riley, V. Humans and Automation: Use, Misuse, Disuse, Abuse. *Hum. Factors* **1997**, *39*, 230–253. <https://doi.org/10.1518/001872097778543886>.
18. Xing, Y.; Lv, C.; Cao, D.; Hang, P. Toward human-vehicle collaboration: Review and perspectives on human-centered collaborative automated driving. *Transp. Res. Part C Emerg. Technol.* **2021**, *128*, 103199. <https://doi.org/10.1016/J.TRC.2021.103199>.
19. Eriksson, A.; Stanton, N.A. Takeover Time in Highly Automated Vehicles: Noncritical Transitions to and from Manual Control. *Hum. Factors* **2017**, *59*, 689–705. <https://doi.org/10.1177/0018720816685832>.
20. Petermeijer, S.; Cieler, S.; Winter, J.d. Comparing spatially static and dynamic vibrotactile take-over requests in the driver seat. *Accid. Anal. Prev.* **2017**, *99*, 218–227. <https://doi.org/10.1016/j.aap.2016.12.001>.
21. Tan, X.; Zhang, Y. When there is no rush to take over: Evaluating multi-stage takeover request designs in conditionally automated driving. *Transp. Res. Part F Traffic Psychol. Behav.* **2026**, *117*, 103465. <https://doi.org/10.1016/j.trf.2025.103465>.
22. George, M.W.; Hass, Z.; Pitts, B.J. Do you trust what an automated vehicle shows you? The effects of presenting dynamic system certainty information on driver behavior. *Appl. Ergon.* **2026**, *134*, 104719. <https://doi.org/10.1016/j.apergo.2025.104719>.
23. Ryu, G.; Ji, Y.G. More explicit is not always better: Boundary conditions for action guidance in hazard notifications across traffic complexity. *Accid. Anal. Prev.* **2026**, *231*, 108510. <https://doi.org/10.1016/j.aap.2026.108510>.
24. Xu, W.; Gao, Z.; Dainoff, M.J. An HCAI methodological framework: Putting it into action to enable human-centered AI. *IEEE Trans. Hum.-Mach. Syst.* **2026**, *56*, 78–94. <https://doi.org/10.1109/THMS.2025.3631590>.
25. Sourav, M.S.G.; Javaid, A.; Cheng, L. A review of AI in human-machine cooperation: Machine perspective. *ACM Trans. Auton. Adapt. Syst.* **2025**, *21*, 1–43. <https://doi.org/10.1145/3774318>.

26. Wu, H.; Zhou, X.; Lyu, N.; Wang, Y.; Xu, L.; Yang, Z. A Review of Methods for Predicting Driver Take-Over Time in Conditionally Automated Driving. *Sensors* **2025**, *25*, 6931. <https://doi.org/10.3390/s25226931>.
27. Davoli, L.; Martalò, M.; Cilfone, A.; Belli, L.; Ferrari, G.; Presta, R.; Montanari, R.; Mengoni, M.; Giralaldi, L.; Amparore, E.; et al. On driver behavior recognition for increased safety: A roadmap. *Safety* **2020**, *6*, 55. <https://doi.org/10.3390/safety6040055>.
28. Uccello, L.; Mastinu, G.; Gobbi, M.; Zaffaroni, I. From Attention to Distraction: Driver State Analysis From an Engineering Perspective—A Review. *IEEE Access* **2025**, *13*, 203286–203305. <https://doi.org/10.1109/ACCESS.2025.3638622>.
29. Wu, Y.; Yao, X.; Deng, F.; Yuan, X. Investigating the impact of interaction design characteristics on trust in automated vehicles. *Transp. Res. Part F Traffic Psychol. Behav.* **2025**, *115*, 103352. <https://doi.org/10.1016/j.trf.2025.103352>.
30. Lee, J.; Abe, G.; Sato, K.; Itoh, M. Developing human-machine trust: Impacts of prior instruction and automation failure on driver trust in partially automated vehicles. *Transp. Res. Part F Traffic Psychol. Behav.* **2021**, *81*, 384–395. <https://doi.org/10.1016/j.trf.2021.06.013>.
31. Deng, M.; Chen, J.; Wu, Y.; Ma, S.; Li, H.; Yang, Z.; Shen, Y. Using voice recognition to measure trust during interactions with automated vehicles. *Appl. Ergon.* **2023**, *116*, 104184. <https://doi.org/10.1016/j.apergo.2023.104184>.
32. Omeiza, D.; Bhattacharyya, R.; Jirotko, M.; Hawes, N.; Kunze, L. A transparency paradox? Investigating the impact of explanation specificity and autonomous vehicle imperfect detection capabilities on passengers. *Transp. Res. Part F Traffic Psychol. Behav.* **2025**, *109*, 1275–1292. <https://doi.org/10.1016/j.trf.2025.01.015>.
33. Peintner, J.; Manger, C.; Berger, M.; Riener, A. Balancing transparency and control: The impact of AI explanation detail on user perception in automated vehicles. *Transp. Res. Interdiscip. Perspect.* **2025**, *34*, 101742. <https://doi.org/10.1016/j.trip.2025.101742>.
34. Irawati, D.A.; Bölükbaşı, E.; Gerber, M.A.; Riener, A. The role of Explainable AI in the Design of Visual Texts for Trust Calibration in Level 3 Automated Vehicles. In *Proceedings of the Adjunct Proceedings of the 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications; AutomotiveUI '24 Adjunct*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 300–303. <https://doi.org/10.1145/3641308.3680514>.
35. Markelius, A.; Lou, Y.; Galazka, M.; Lundgren, S.; Zemblys, R.; Lind, H.; Lowe, R. Investigating the Mitigation of Stress in Autonomous and Non-autonomous Vehicles Using LLM Feedback. In *Proceedings of the Interacción*; Springer Nature: Cham, Switzerland, 2025; pp. 108–127. https://doi.org/10.1007/978-3-031-92692-1_7.
36. Naujoks, F.; Purucker, C.; Neukum, A. Secondary task engagement and vehicle automation—Comparing the effects of different automation levels in an on-road experiment. *Transp. Res. Part F Traffic Psychol. Behav.* **2016**, *38*, 67–82. <https://doi.org/10.1016/J.TRF.2016.01.011>.
37. Merlhiot, G.; Bueno, M. How drowsiness and distraction can interfere with take-over performance: A systematic and meta-analysis review. *Accid. Anal. Prev.* **2021**, *170*, 106536. <https://doi.org/10.1016/j.aap.2021.106536>.
38. Wischniewski, M.; Krämer, N.; Müller, E. Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. In *Proceedings of the International Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2023; pp. 1–16. <https://doi.org/10.1145/3544548.3581197>.
39. Louw, T.; Merat, N. Are you in the loop? Using gaze dispersion to understand driver visual attention during vehicle automation. *Transp. Res. Part C Emerg. Technol.* **2017**, *76*, 35–50. <https://doi.org/10.1016/j.trc.2017.01.001>.
40. Mackworth, N.H. The breakdown of vigilance during prolonged visual search. *Q. J. Exp. Psychol.* **1948**, *1*, 6–21. <https://doi.org/10.1080/17470214808416738>.
41. Hebb, D.O. Drives and the CNS (conceptual nervous system). *Brain Physiol. Psychol.* **2023**, *62*, 67–83. <https://doi.org/10.2307/jj.8501509.7>.
42. Nieuwenhuis, S. Arousal and performance: Revisiting the famous inverted-U-shaped curve. *Trends Cogn. Sci.* **2024**, *28*, 394–396. <https://doi.org/10.1016/j.tics.2024.03.011>.
43. Endsley, M.R. Toward a Theory of Situation Awareness in Dynamic Systems. *Hum. Factors* **1995**, *37*, 32–64. <https://doi.org/10.1518/001872095779049543>.
44. Kaplan, S.; Guvensan, M.A.; Yavuz, A.G.; Karalurt, Y. Driver Behavior Analysis for Safe Driving: A Survey. *IEEE Trans. Intell. Transp. Syst.* **2015**, *16*, 3017–3032. <https://doi.org/10.1109/TITS.2015.2462084>.
45. Dong, Y.; Hu, Z.; Uchimura, K.; Murayama, N. Driver inattention monitoring system for intelligent vehicles: A review. *IEEE Trans. Intell. Transp. Syst.* **2009**, *12*, 875–880. <https://doi.org/10.1109/TITS.2010.2092770>.
46. Vass, S.; Donà, R.; Mattas, K.; Rövid, A.; Vincze, Z.; Schöner, H.P.; Csonthó, M.; Pataki, M.; Galassi, M.C.; Tihanyi, V.; et al. *Public Road Tests of Automated Lane Keeping Systems Under UN Regulation No. 157*; Technical report; UNECE: Geneva, Switzerland, 2026. https://doi.org/10.1007/978-3-031-88974-5_9.
47. Zhu, E.; Yang, S. Towards human digital twin: Reviewing human modelling and simulation. *J. Ind. Inf. Integr.* **2025**, *48*, 100975. <https://doi.org/10.1016/j.jii.2025.100975>.
48. Politis, I.; Brewster, S.; Pollick, F. Language-Based Multimodal Displays for the Handover of Control in Autonomous Cars. In *Proceedings of the International Conference on Automotive User Interfaces and Interactive Vehicular Applications*; ACM: New York, NY, USA, 2015; pp. 3–10. <https://doi.org/10.1145/2799250.2799262>.

49. Presta, R.; Simone, F.D.; Tancredi, C.; Chiesa, S. Nudging the safe zone: Design and assessment of HMI strategies based on intelligent driver state monitoring systems. In *Proceedings of the Interacción*; Springer Nature: Cham, Switzerland, 2023; pp. 166–185. https://doi.org/10.1007/978-3-031-35678-0_10.
50. Ceccacci, S. Designing in-car emotion-aware automation. *Eur. Transp. Eur.* **2021**, *84*, 1–15. <https://doi.org/10.48295/et.2021.84.5>.
51. Braun, M.; Weber, F.; Alt, F. Affective Automotive User Interfaces—Reviewing the State of Driver Affect Research and Emotion Regulation in the Car. *ACM Comput. Surv.* **2020**, *54*. <https://doi.org/10.1145/3460938>.
52. Presta, R.; Tancredi, C.; Simone, F.D.; Iacono, M.; Mancuso, L. Improving time to take over through HMI strategies nudging a safe driving state. In *Proceedings of the Interacción*; Springer Nature: Cham, Switzerland, 2024; pp. 27–43. https://doi.org/10.1007/978-3-031-60477-5_3.
53. Lee, J.D.; See, K.A. Trust in Automation: Designing for Appropriate Reliance. *Hum. Factors* **2004**, *46*, 50–80. <https://doi.org/10.1518/hfes.46.1.50.30392>.
54. Hoff, K.A.; Bashir, M. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Hum. Factors* **2015**, *57*, 407–434. <https://doi.org/10.1177/0018720814547570>.
55. Zhang, R.; Tan, Z.; Lin, Z.; Zhang, R.; Liu, C. Exploring the trust and behavior of experienced advanced driver assistance system drivers: An on-road study. *Accid. Anal. Prev.* **2025**, *217*, 108071. <https://doi.org/10.1016/j.aap.2025.108071>.
56. Körber, M. Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. *Adv. Intell. Syst. Comput.* **2018**, *823*, 13–30. https://doi.org/10.1007/978-3-319-96074-6_2.
57. Lee, M.; Pitts, B.J. Sometimes right, sometimes wrong: Drivers’ responses to inconsistently accurate automated vehicle system confidence information. *Accid. Anal. Prev.* **2026**, *228*, 108412. <https://doi.org/10.1016/j.aap.2026.108412>.
58. Kim, J.; Rohrbach, A.; Darrell, T.; Canny, J.; Akata, Z. Textual explanations for self-driving vehicles. In *Proceedings of the European Conference on Computer Vision*; Springer International Publishing: Cham, Switzerland, 2018; pp. 577–593. https://doi.org/10.1007/978-3-030-01216-8_35.
59. Zablocki, É.; Ben-Younes, H.; Pérez, P.; Cord, M. Explainability of deep vision-based autonomous driving systems: Review and challenges. *Int. J. Comput. Vis.* **2022**, *130*, 2425–2452. <https://doi.org/10.1007/s11263-022-01657-x>.
60. Atakishiyev, S.; Salameh, M.; Yao, H.; Goebel, R. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. *IEEE Access* **2024**, *12*, 101603–101625. <https://doi.org/10.1109/ACCESS.2024.3431437>.
61. Shneiderman, B. Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. *ACM Trans. Interact. Intell. Syst.* **2020**, *10*, 1–31. <https://doi.org/10.1145/3419764>.
62. Chander, B.; John, C.; Warriar, L.; Gopalakrishnan, K. Toward trustworthy artificial intelligence (TAI) in the context of explainability and robustness. *ACM Comput. Surv.* **2024**, *57*, 1–49. <https://doi.org/10.1145/3675392>.
63. Zang, J.; Jeon, M. When more is less: Finding the optimal balance of intelligent agents’ transparency in level 3 automated vehicles. *Int. J. Hum. Comput. Stud.* **2024**, *193*, 103384. <https://doi.org/10.1016/j.ijhcs.2024.103384>.
64. Cui, C.; Ma, Y.; Cao, X.; Ye, W.; Wang, Z. Drive as You Speak: Enabling Human-Like Interaction with Large Language Models in Autonomous Vehicles. In *Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, Waikoloa, HI, USA, 1–6 January 2024; pp. 384–393. <https://doi.org/10.1109/WACVW60836.2024.00101>.
65. Shao, H.; Hu, Y.; Wang, L.; Waslander, S.L.; Liu, Y.; Li, H. LMDrive: Closed-Loop End-to-End Driving with Large Language Models. In *Proceedings of the Computer Vision and Pattern Recognition*, Seattle, WA, USA, 16–22 June 2024; pp. 15120–15130. <https://doi.org/10.1109/CVPR52733.2024.01432>.
66. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.; Chen, D.; Dai, W.; et al. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2022**, *55*, 248:1–248:38. <https://doi.org/10.1145/3571730>.
67. Lin, Y. Toward Intelligent Human Machine Interactions. *Mech. Eng.* **2017**, *139*, S4–S8. <https://doi.org/10.1115/1.2017-jun-4>.
68. Cai, H.; Lin, Y. Coordinating Cognitive Assistance With Cognitive Engagement Control Approaches in Human–Machine Collaboration. *IEEE Trans. Syst. Man. Cybern. Part A Syst. Hum.* **2012**, *42*, 286–294. <https://doi.org/10.1109/TSMCA.2011.2169953>.
69. Yang, G.; Lin, Y.; Bhattacharya, P. A Driver Fatigue Recognition Model Based on Information Fusion and Dynamic Bayesian Network. *Inf. Sci.* **2010**, *180*, 1942–1954. <https://doi.org/10.1016/j.ins.2010.01.011>.
70. Lin, Y.; Tang, P.; Zhang, W.J.; Yu, Q. Artificial neural network modelling of driver handling behaviour in a driver-vehicle-environment system. *Int. J. Veh. Des.* **2005**, *37*, 24–45. <https://doi.org/10.1504/IJVD.2005.006087>.
71. Hart, S.G.; Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*; Hancock, P.A., Meshkati, N., Eds.; Elsevier: Amsterdam, The Netherlands, 1988; pp. 139–183. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9).
72. Seppelt, B.D.; Lee, J.D. Making Adaptive Cruise Control (ACC) Limits Visible. *Int. J. Hum. Comput. Stud.* **2007**, *65*, 192–205. <https://doi.org/10.1016/j.ijhcs.2006.10.001>.
73. Wang, Z.; Gupta, R.; Han, K.; Wang, H.; Ganlath, A.; Ammar, N.; Tiwari, P. Mobility digital twin: Concept, architecture, case study, and future challenges. *IEEE Internet Things J.* **2022**, *9*, 17452–17467. <https://doi.org/10.1109/JIOT.2022.3156028>.

74. Wang, K.; Li, Z.; Nonomura, K.; Yu, T.; Sakaguchi, K.; Hashash, O.; Saad, W. Smart mobility digital twin based automated vehicle navigation system: A proof of concept. *IEEE Trans. Intell. Veh.* **2024**, *9*, 4348–4361. <https://doi.org/10.1109/TIV.2024.3368109>.
75. Hu, Z.; Lou, S.; Xing, Y.; Wang, X.; Cao, D.; Lv, C. Review and perspectives on driver digital twin and its enabling technologies for intelligent vehicles. *IEEE Trans. Intell. Veh.* **2022**, *7*, 417–440. <https://doi.org/10.1109/TIV.2022.3195635>.
76. Gaffinet, B.; Ali, J.A.H.; Naudet, Y.; Panetto, H. Human digital twins: A systematic literature review and concept disambiguation for industry 5.0. *Comput. Ind.* **2025**, *166*, 104230. <https://doi.org/10.1016/j.compind.2024.104230>.
77. Asad, U.; Khan, M.; Khalid, A.; Lughmani, W.A. Human-centric digital twins in industry: A comprehensive review of enabling technologies and implementation strategies. *Sensors* **2023**, *23*, 3938. <https://doi.org/10.3390/s23083938>.
78. VanDerHorn, E.; Mahadevan, S. Digital Twin: Generalization, Characterization and Implementation. *Decis. Support Syst.* **2021**, *145*, 113524. <https://doi.org/10.1016/j.dss.2021.113524>.
79. Puliafito, A.; Tricomi, G.; Zafeiropoulos, A.; Papavassiliou, S. Smart Cities of the Future as Cyber Physical Systems: Challenges and Enabling Technologies. *Sensors* **2021**, *21*, 3349. <https://doi.org/10.3390/s21103349>.
80. Thonhofer, E.; Sigl, S.; Fischer, M.; Heuer, F.; Kuhn, A.; Erhart, J.; Harrer, M.; Schildorfer, W. Infrastructure-based digital twins for cooperative, connected, automated driving and smart road services. *IEEE Open J. Intell. Transp. Syst.* **2023**, *4*, 311–324. <https://doi.org/10.1109/OJITS.2023.3266800>.
81. Aksjonov, A.; Kyrki, V. A safety-critical decision-making and control framework combining machine-learning-based and rule-based algorithms. *SAE Int. J. Veh. Dyn. Stab. NVH* **2023**, *7*, 287–299. <https://doi.org/10.4271/10-07-03-0018>.
82. Kabir, S.; Papadopoulos, Y. Applications of Bayesian networks and Petri nets in safety, reliability, and risk assessments: A review. *Saf. Sci.* **2019**, *115*, 154–175. <https://doi.org/10.1016/J.SSCI.2019.02.009>.
83. Manas, K.; Paschke, A. Knowledge integration strategies in autonomous vehicle prediction and planning: A comprehensive survey. In *Proceedings of the 2025 IEEE Intelligent Vehicles Symposium (IV)*; IEEE: New York, NY, USA, 2025; pp. 694–701. <https://doi.org/10.1109/IV64158.2025.11097371>.
84. Bhatt, U.; Zhang, Y.; Antorán, J.; Liao, Q.; Sattigeri, P.; Fogliato, R.; Melançon, G.G.; Krishnan, R.; Stanley, J.; Tickoo, O.; et al. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*; AIES '21; Association for Computing Machinery: New York, NY, USA, 2020; pp. 401–413. <https://doi.org/10.1145/3461702.3462571>.
85. Slagel, J.T.; White, L.M.; Dutle, A.; Muñoz, C.A.; Crespo, N. *A Verification Framework for Runtime Assurance of Autonomous UAS*; Technical Report NASA/TM-20240007986; NASA Langley Research Center: Hampton, VA, USA, 2024. <https://doi.org/10.1109/dasc62030.2024.10748654>.
86. Mehmood, U.; Sheikhi, S.; Bak, S.; Smolka, S.A.; Stoller, S.D. The black-box simplex architecture for runtime assurance of autonomous CPS. In *Proceedings of the NASA Formal Methods*; Springer: Cham, Switzerland, 2021; pp. 231–250. https://doi.org/10.1007/978-3-031-06773-0_12.
87. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689, 12 July 2024, 2024. Available online: <http://data.europa.eu/eli/reg/2024/1689/oj> (accessed on 5 June 2026).
88. Naujoks, F.; Wiedemann, K.; Schömig, N.; Jarosch, O.; Gold, C. Expert-based controllability assessment of control transitions from automated to manual driving. *MethodsX* **2018**, *5*, 579–592. <https://doi.org/10.1016/j.mex.2018.05.007>.
89. Hu, X.; Li, S.; Huang, T.; Tang, B.; Huai, R.; Chen, L. How simulation helps autonomous driving: A survey of sim2real, digital twins, and parallel intelligence. *IEEE Trans. Intell. Veh.* **2023**, *9*, 593–612. <https://doi.org/10.1109/TIV.2023.3312777>.
90. Schwindt-Drews, S.; Abendroth, B. Mental model evolvment during drivers' first experience with conditionally automated driving systems in real-world traffic. *Transp. Res. Part F Traffic Psychol. Behav.* **2026**, *118*, 103508. <https://doi.org/10.1016/j.trf.2026.103508>.
91. Kim, S.; Oviedo-Trespalacios, O. Behavioural adaptation in advanced driver assistance systems (ADAS) use. *Appl. Ergon.* **2026**, *136*, 104799. <https://doi.org/10.1016/j.apergo.2026.104799>.
92. Wang, A.; Wang, J.; Shi, W.; He, D. Cognitive Workload Estimation in Conditionally Automated Vehicles Using Transformer Networks Based on Physiological Signals. *Transp. Res. Rec.* **2024**, *2678*, 1183–1196. <https://doi.org/10.1177/03611981241250023>.
93. Nilsson, E.J.; Bärghman, J.; Aust, M.L.; Matthews, G.; Svanberg, B. Let Complexity Bring Clarity: A Multidimensional Assessment of Cognitive Load Using Physiological Measures. *Front. Neuroergonomics* **2022**, *3*, 787295. <https://doi.org/10.3389/fnrgo.2022.787295>.
94. AlArnaout, Z.; Zaki, C.; Kotb, Y.; AlAkkoumi, M.; Mostafa, N. Exploiting heart rate variability for driver drowsiness detection using wearable sensors and machine learning. *Sci. Rep.* **2025**, *15*, 10542. <https://doi.org/10.1038/s41598-025-08582-2>.
95. Parasuraman, R.; Molloy, R.; Singh, I.L. Performance Consequences of Automation-Induced 'Complacency'. *Int. J. Aviat. Psychol.* **1993**, *3*, 1–23. https://doi.org/10.1207/s15327108ijap0301_1.
96. Parasuraman, R.; Sheridan, T.B.; Wickens, C.D. A Model for Types and Levels of Human Interaction with Automation. *IEEE Trans. Syst. Man. Cybern. Part A* **2000**, *30*, 286–297. <https://doi.org/10.1109/3468.844354>.

97. Hancock, P.; Billings, D.; Schaefer, K.E.; Chen, J.Y.; Visser, E.D.; Parasuraman, R. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Hum. Factors* **2011**, *53*, 517–527. <https://doi.org/10.1177/0018720811417254>.
98. Waytz, A.; Heafner, J.; Epley, N. The Mind in the Machine: Anthropomorphism Increases Trust in an Autonomous Vehicle. *J. Exp. Soc. Psychol.* **2014**, *52*, 113–117. <https://doi.org/10.1016/j.jesp.2014.01.005>.
99. Stampf, A.; Colley, M.; Girst, B.; Rukzio, E. Exploring Passenger-Automated Vehicle Negotiation Utilizing Large Language Models for Natural Interaction. In *Proceedings of the International Conference on Automotive User Interfaces and Interactive Vehicular Applications*; ACM: New York, NY, USA, 2024; pp. 350–362. <https://doi.org/10.1145/3640792.3675725>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.