# LLM-assisted Systematic Review of Inaccurate Mental Models for Driving Automation Systems: Constructs, KPIs and Measurements

YING FANG, Eindhoven University of Technology, The Netherlands
PAVLO BAZILINSKYY, Eindhoven University of Technology, The Netherlands
MARIEKE MARTENS, Eindhoven University of Technology, The Netherlands

Research on the interaction between users and driving automation systems (DAS) is growing rapidly. To support safe interaction, researchers have examined how drivers comprehend these systems, yielding the concept of mental models. The literature shows that users naturally form incomplete representations, termed inaccurate mental models (IMM). However, research on measuring IMM remains fragmented: studies typically isolate single dimensions, relying on declarative knowledge questionnaires, behavioural indicators, or related constructs such as miscalibrated trust, and no coherent framework links these metrics to the cognitive mechanisms underlying safety-critical risks. In this systematic review, we investigate how IMM is interpreted and assessed in in-vehicle human–DAS interaction at SAE Levels 1–4, with the goal of developing a unified assessment framework that supports systematic diagnosis. Methodologically, we used a large language model (GPT-5.2) as an independent second reviewer working in parallel with a human expert, introducing a collaborative human–LLM screening strategy for human–machine interface review research. Of 1,178 unique records, 53 were included (38 empirical and 15 theoretical). We organise the evidence into an initial mental model (explicit and implicit conceptions) and a situational mental model (situational understanding, situational attitude, and behaviour). Across these five dimensions, we identify 7 constructs and 30 key performance indicator (KPI) categories and map them to measurement methods, automation levels, and driving contexts. The framework supports developers and researchers in choosing complementary measures and converting abstract cognitive risks into quantifiable data, with the aim of diminishing inaccurate mental models for safe human-machine interaction (HMI).

Additional Key Words and Phrases: Inaccurate mental model, human-automation interaction, driving automation system, assessment framework, systematic review, large language model

## 1 Introduction

Humans do not always require a precise understanding of technology to be able to interact with it. Often, technology is so intricate that it cannot be fully grasped by its users, so people instead depend on an internal model that captures only the most important elements—one that is naturally incomplete (Norman, 1983b). The notion of a mental model emerged to account for this meta-cognitive process: an internal representation of the world to test potential actions and anticipate their consequences before carrying them out in the physical environment (Craik, 1967, Johnson-Laird, 1983, Moore and Golledge, 1976). Within this tradition, mental models are understood as dynamic inferential frameworks (Collins and Gentner, 1987, Gentner, 1983, Johnson-Laird, 1983) and become central to the domain of human–automation interaction, where users must cope with increasingly opaque digital systems. We adopt the definition of a mental model as an evolvable rather than static mechanism (Beggiato and Krems, 2013, Beggiato et al., 2015, Hegarty et al., 2013), through

Authors' Contact Information: Ying Fang, Eindhoven University of Technology, Eindhoven, The Netherlands, y.fang1@tue.nl; Pavlo Bazilinskyy, Eindhoven University of Technology, Eindhoven, The Netherlands, p.bazilinskyy@tue.nl; Marieke Martens, Eindhoven University of Technology, Eindhoven, The Netherlands, m.h.martens@tue.nl.


Manuscript submitted to ACM

which a user explains, interprets, and predicts (Johnson-Laird, 1983, Norman, 1983a) the purpose, functioning, and state of a complex automation system (Parasuraman and Riley, 1997, Parasuraman et al., 2000).

Building an adequate mental model is fundamental for safety when people use driving automation systems (DAS) (Beggiato and Krems, 2013, Qiao et al., 2024). Nonetheless, establishing such a model has always been challenging, particularly in the context of automated driving. Contemporary vehicles span a spectrum of automation as defined by the Society of Automotive Engineers (SAE International, 2021), from Level 0 (fully manual) to Level 5 (fully automated), with this spectrum filling rapidly but unevenly. Mercedes-Benz and BMW launched certified Level 3 conditional automated driving (CAD) in 2022 and 2024 in Germany (BMW Group, 2023, Mercedes-Benz Group, 2022), respectively, enabling drivers to legally disengage from the driving task within specified operational design domains. By contrast, Tesla's Full Self-Driving (Supervised), fundamentally a Level 2 driver assistance feature (Council, 2026), received provisional European type approval from the RDW in April 2026 (RDW, 2026), allowing immediate deployment in the Netherlands while EU-wide approval is still pending. Although all are nominally positioned on the same SAE scale, these systems differ markedly in naming, capability, and the moment-to-moment demands they place on the user. Consequently, these largely technology- and regulation-oriented levels may not reflect how users actually understand control authority (Tabone et al., 2021, Yang et al., 2017). Recent real-world crashes show some misunderstanding can be fatal, as in the Xiaomi SU7 collision involving its Level 2 motorway assistance system (Reuters, 2025), which regulators and analysts warn may cause users to misinterpret its function and be unready to intervene when required.

We use **inaccurate mental model (IMM)** to describe incomplete or erroneous representations of system capability and responsibility (Badke-Schaub et al., 2007, Harari et al., 2019, Kwon et al., 2016, Santos et al., 2016). We argue that such inaccuracies should not be seen simply as user failings, as if users ought to grasp every technical nuance of ever more complex automation. Instead, IMM should be interpreted as a diagnostic indicator of how effectively the automated system supports the interaction and sufficiently conveys its status, capabilities, and limitations, thereby enabling safe interaction (Carsten and Martens, 2019, de Winter et al., 2023, Tinga et al., 2023). Therefore, the primary objective of automated driving research is to enhance HMI quality so that it reliably closes the gap between machine logic and human expectations, ultimately enabling safe automated vehicles.

However, research on assessing **IMM** is fragmented into three primary strands, each marked by distinct structural shortcomings: (i) The first strand defines mental model accuracy along a single dimension only, most often through standardised questionnaires that assess declarative knowledge about how the system operates (Beggiato et al., 2015, Boelhouwer et al., 2019, Seupke et al., 2025a). These instruments assess what the system can do, the conditions under which it operates, and the aspects for which the driver remains responsible, and are often tailored to specific system features under SAE Levels 1–4. (ii) the second line represents IMM derived from behavioural indicators, such as observable patterns of misuse, disuse, or unjustified reliance on automation (Banks et al., 2018, Parasuraman and Riley, 1997, Wörle and Metz, 2023). (iii) A third stream approaches IMM through related constructs, such as miscalibrated trust (Hoff and Bashir, 2015, Lee and See, 2004), mode confusion (Eom and Lee, 2015, 2022), or diminished situation awareness (de Winter et al., 2014, Endsley, 1995), and treats these phenomena as downstream indicators from which inaccuracies in mental models can be inferred. However, each strand, taken alone, captures only a partial slice of the construct and leaves systematic blind spots. It is important to recognise that a user's mental model is neither a fixed measure of knowledge nor just one specific behaviour or attitude. What people claim does not always align with how they actually behave: users who state they understand these limits may still misjudge what is required of them when they must intervene in time-critical scenarios (Boelhouwer et al., 2019), or still exhibit over-reliance on the system (Victor et al., 2018) even while maintaining "eyes on the road and hands on the wheel". A mental model is

a dynamic process linking perception (Novakazi, 2025), interpretation, and action across contexts, spanning pre-use beliefs, in-the-moment sense-making and the way experiences reshape understanding. Therefore, a knowledge gap remains: without a shared framework for fully evaluating IMM, researchers are left with a piecemeal toolkit, inferring how isolated measures map onto the cognitive mechanisms underpinning safety-critical issues.

### 1.1 Aim of study

To resolve the existing fragmentation, this paper presents a systematic review of how IMM is conceptualised and assessed. We focus specifically on the internal perspective of the driver–automation relationship from SAE Levels 1–4, encompassing a shifting allocation of control: from continuous driver assistance (e.g., ACC, LKA at SAE Level 1/2), to conditional automation requiring fallback readiness (SAE Level 3), up to highly automated driving (SAE Level 4)[1]. Our core objective is to move beyond isolated metrics by identifying exactly which theoretical **constructs** and **key performance indicators (KPIs)** genuinely signal an IMM. This architecture explicitly maps underlying cognitive constructs to their corresponding subjective and objective KPIs. By synthesising these elements, we provide researchers and practitioners with the precise diagnostic tools necessary to evaluate automotive HMI design in the aforementioned context, translating abstract cognitive risks into measurable and actionable data.

To guide this synthesis, the review addresses the following overarching research questions:

- **RQ1: Conceptualise IMM.** Which constructs are employed to characterise the IMM in human–DAS interaction, and how are these constructs defined and differentiated across the various studies?
- **RQ2: Operationalise and apply IMM assessment.** Which KPIs and assessment methods are used to make IMM-related constructs measurable (e.g., questionnaires/interviews, probes/tasks, logs/telemetry, eye tracking/physiological signals), and how are these combinations distributed across automation levels, driving functions, and traffic scenarios?

### 1.2 Conceptual framework

To systematically map this cognitive landscape, we firstly propose a two-layer mental model framework for concept distinction (Figure 1): (i) the **initial mental model** is viewed as stable knowledge structures in long-term memory (Craik, 1967) that persist across situations (Gentner, 1983, Jones et al., 2011, Rouse and Morris, 1986), which in context of automated driving, represent users' prior beliefs, experiences, and knowledge about how the automated system works before the actual interaction (Carsten and Martens, 2019, Novakazi et al., 2025, Seupke et al., 2025a, Walker et al., 2023); (ii) the **situational mental model** is seen as short-lived constructions in working memory, built to reason about a particular unfolding situation (Durso et al., 2007, Endsley, 2015, Johnson-Laird, 1983, Kintsch, 1998).

Our framework additionally distinguished five dimensions to specify constructs and measurement approaches in the context of human-DAS interaction. We anchor the *initial mental model* in Nelson and Narens' meta-memory framework (Nelson, 1990), resulting in two dimensions. *D1 Explicit conception* represents an object level, encompassing the user's declarative knowledge (Anderson, 2013) regarding the system features (its capabilities, limitations, and mode logic) (Seupke et al., 2025a). *D2 Implicit conception* represents a meta-level, that is, what users believe they understand and how confident they feel about that understanding, regardless of whether this feeling is correct (Greenwood et al., 2022, Metz et al., 2024). Thus, a user may feel very confident while harboring substantial misconceptions (Sanbonmatsu et al., 2018), or feel uncertain despite in fact possessing an essentially adequate model. The *situational mental model* is

[1]Here, SAE Level 4 refers specifically to driver-facing dual-mode vehicles rather than pure Level 4 systems operating without any human intervention.

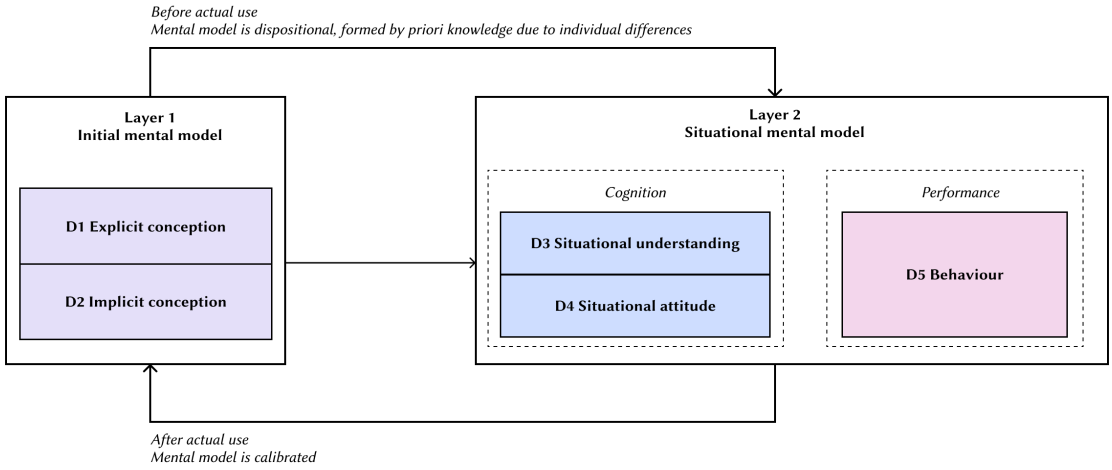


Fig. 1. Conceptual Framework of the development of mental model in human-AV interaction

the live cognitive structure when users actively use DAS, where prior knowledge combines with real-time interface feedback and vehicle behaviour, shifting them from 'espoused theory' (what they say they know) to 'theory in use' (how they actually interpret and act). We capture this theory in use through the framework's final three dimensions: *D3 Situational understanding* concerns whether the user correctly grasps the vehicle's immediate intent and can anticipate its next actions in a given scenario (Novakazi, 2025). *D4 Situational attitude* captures momentary trust and acceptance in the specific driving context (Seupke et al., 2025b, Walker et al., 2023). It is a short-lived, context-bound judgement of how satisfied the user is and how willing they are to rely on the system for the current hazard. *D5 Behaviour* comprises observable actions that reveal this internal cognitive state, indicating whether the driver's "theory in use" supports safe, well-coordinated interaction with the vehicle.

## 2 Method

We conducted the systematic review with PRISMA 2020 guideline (Page et al., 2021). In parallel with human screening, we integrated the Large Language Model as a collaborative second reviewer to improve the rigour and efficiency of study selection. Study identification involved: (i) a database search and (ii) a collaborative, dual-reviewer screening procedure in which a human reviewer and a Large Language Model (GPT-5.2) [2] independently applied the same eligibility criteria during both title, abstract and full-text screening. The study selection procedure is illustrated in Figure 2.

### 2.1 Database search

Four bibliographic databases were selected: ACM Digital Library (https://dl.acm.org), IEEE Xplore (https://ieeexplore.ieee.org), Scopus (https://www.scopus.com), and Web of Science (https://www.webofscience.com). ACM and IEEE were chosen for their strong coverage of human-computer interaction and automotive research, including key venues in driving automation and in-vehicle interface design. Scopus was included to broaden coverage across interdisciplinary

[2]GPT-5.2 was selected as the highest-performing general-purpose LLM available at the time of screening (January 2026 ). We acknowledge that LLM capabilities evolve rapidly; this choice reflects the state of the art during data collection rather than an endorsement of a specific vendor.

Table 1. General query structure for all searches

| Context | Keywords |
|---|---|
| Automated Driving | "automated driving" OR "autonomous driving" OR "automated vehicle" OR "operatorless" OR "self-driving" |
| Mental Model | "mental model" OR awareness OR understanding |
| Interaction | "HMI" OR "human–automation interaction" OR "human–vehicle interaction" OR "human–machine interaction" |
| Safe | Safe |
| Negative Factors | "error" OR "false" OR "misalignment" OR "mismatch" OR "incorrect" OR "inaccurate" OR "confusion" OR "surprise" |

journals and conference proceedings, while Web of Science provided complementary indexing and an additional cross-check for retrieval completeness.

*2.1.1 Search queries and filters.* To identify relevant publications, we formulated database-specific search strings using Boolean operators. The search query comprised keywords from five conceptual blocks, which were searched separately: automated driving, mental models/understanding, interaction/HMI, and terms representing negative factors. The structure of the query and the keywords used in each block are presented in Table 1.

Additionally, three filters were applied consistently across databases to improve relevance: (i) **Publication year**. We limited the search to studies published between 2000 and 2025 to capture both foundational work on mental models in human–automation interaction and more recent evidence reflecting contemporary automated-driving technologies and in-vehicle HMI developments; (ii) **Publication type** restricted to peer-reviewed journal articles and conference papers; (iii) **Language** restricted to English.

*2.1.2 Search results.* Using the predefined search queries (Table 1), the database search identified a total of 1,541 records (ACM: 546; IEEE: 489; Scopus: 467; Web of Science: 39). The database search was executed on 06 November 2025, on which date all retrieved records were exported, combined into a single dataset, and stored as a CSV file. Since individual publications may appear in multiple databases, duplicate entries and clearly ineligible records were subsequently removed using a Python-based deduplication workflow, leaving **1,178** unique records for title and abstract screening after the exclusion of 363 records (49 duplicates; 314 proceedings front-matter records).

## 2.2 Eligibility criteria

We defined eligibility criteria as *a priori* to ensure that included studies explicitly addressed IMM in the context of **in-vehicle human–automation interaction** in automated driving, and that evidence was grounded in human factors data. The criteria were developed to align with the study aim and research questions, and were operationalised as three inclusion criteria (IC1-IC3) and six exclusion criteria (EC1-EC6).

*2.2.1 Inclusion criteria.* Studies were included if they met all of the following:

- **IC1 Direct interpretation or measurements of mental model.** The title/abstract explicitly framed the work as mental-model–relevant in human–AV interaction (e.g., understanding/interpretation, (mis)perception, misunderstanding, expectation mismatch, mode confusion, automation surprise, or misuse/disuse attributable to incorrect understanding).

- **IC2 Relevance of mental model.** The study employed constructs or measures that can diagnose users' inferences about system capability or state (i.e., mental model content), and interpreted outcomes as evidence of users' understanding or belief. Eligible measures included cognitive/interpretive indicators (e.g., system comprehension, situation awareness when framed as correctness regarding ADS state/limits, perceived capability, explicit knowledge/probe tests, prediction tasks), behavioural/performance indicators (e.g., mode errors, takeover errors, inappropriate reliance/misuse/disuse, intervention timing, incorrect actions relative to system limits), and trust/workload/attention measures only when explicitly interpreted as reflecting incorrect beliefs/inferences or miscalibrated reliance (rather than generic workload or attention differences).
- **IC3 Review/Conceptual eligible.** Include if the paper defines/argues mechanisms of inaccurate understanding (complacency, mode awareness, misuse/disuse pitfalls) or synthesises evaluation methods relevant to diagnosing user understanding.

*2.2.2 Exclusion criteria.* Studies were excluded if any of the following applied:

- **EC1 Not peer-reviewed.** Not peer-reviewed or not accessible (e.g., white papers, posters, or non-peer-reviewed preprints).
- **EC2 Not in context of automated driving.** Studies outside SAE J3016-defined automotive driving automation (e.g., maritime, aviation), and studies of manual driving or SAE Level 0 vehicles in which no sustained driving automation is present (since IMM as conceptualised here requires the existence of automation capability to be misrepresented).
- **EC3 Absence of human factors.** System-only contributions without human participants, human data, or human-in-the-loop evaluation, and without cognitive/behavioural/perceptual/subjective measures relevant to users.
- **EC4 UX-only or non-critical without safety link.** Primarily comfort, acceptance, preference, aesthetics, or general UX in routine driving without a safety-relevant link to misalignment (e.g., misuse/disuse, delayed reactions, or other safety-critical behaviour).
- **EC5 Non-English publication.** Not written in English.
- **EC6 Out-of-scope interaction types.** Exterior eHMI/pedestrian or other road-user interaction; operators outside the vehicle; remote operators in L4; passenger-only interaction; other road users(ORU)-AV interaction.

## 2.3 Screening

The study selection followed a two-stage screening process comprising: (i) **title and abstract screening** and (ii) **full-text screening**, based on PRISMA 2020 recommendations for transparent reporting of selection procedures (Page et al., 2021). The overall screening workflow and the corresponding record numbers are presented in Figure 2. The screening was conducted in parallel by two independent reviewers (human and LLM), and any disagreements were addressed using a structured resolution procedure. At every stage, both reviewers independently applied the same eligibility criteria (IC1-IC3; EC1-EC6).

*2.3.1 Manual title and abstract screening.* The first author screened all remaining records in the title and abstract screening process against the eligibility criteria (see Section 2.2). Of the 1,178 records screened, **52** were judged eligible for full-text assessment and 1,126 were excluded at this stage (Figure 2).

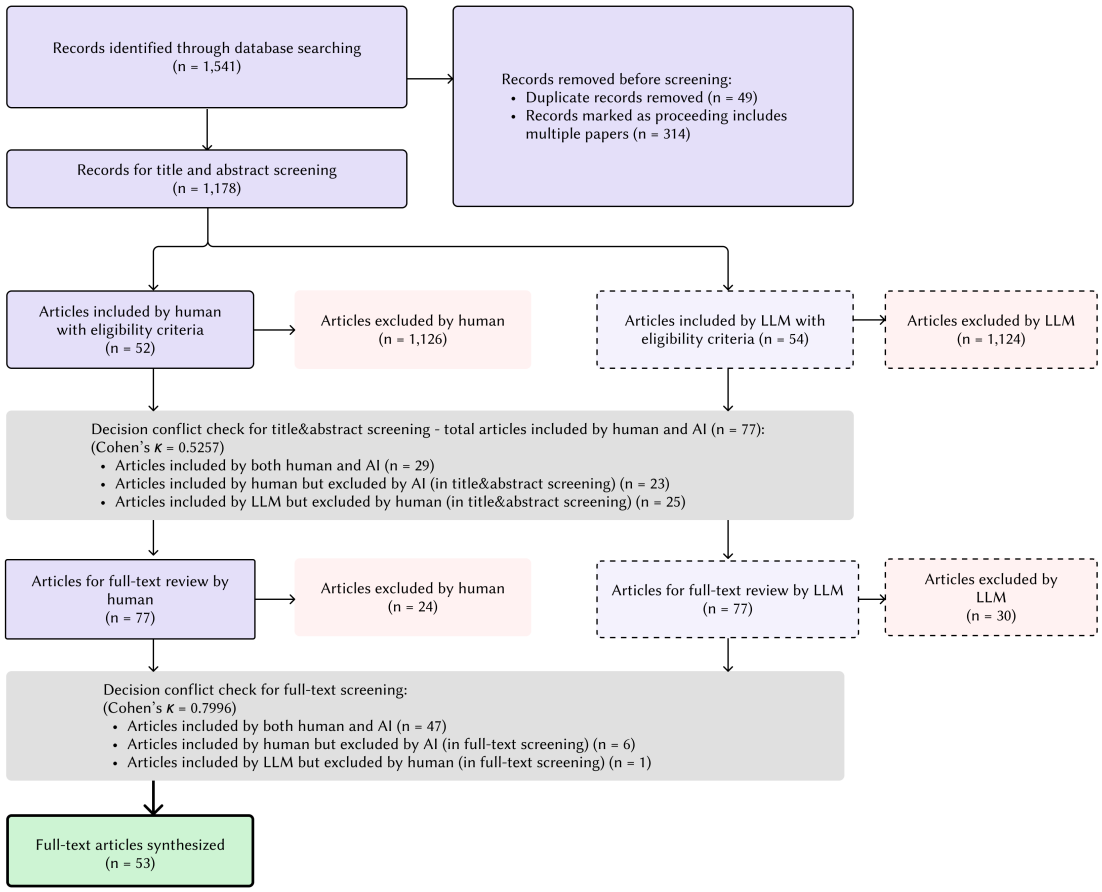


Fig. 2. A structured process of review illustrating study selection and the human-LLM dual-reviewer process.

*2.3.2 LLM title and abstract screening.* Prior work has shown that LLMs can effectively support systematic review (Borges et al., 2026, Chen et al., 2026), conducting title and abstract screening (Dennstädt et al., 2024, Zhuang et al., 2025). In this study, we used an LLM as a second independent reviewer for two main reasons: (i) to **mitigate human error and maximise recall**. Evidence shows that even with two human reviewers resolving conflicts, about 3% of relevant citations are missed; this rises to 13% with a single reviewer (Gartlehner et al., 2020). Serving as a non-fatigable verification layer, the LLM increases sensitivity. Recent trials suggest that although LLMs may have lower precision than humans, they can attain near-perfect recall (96-100%) in binary screening tasks (Guo et al., 2024), functioning as a "safety net" for relevant papers that fatigued human reviewers may overlook. (ii) to **address a methodological gap in the specialised domain of Human-Machine Interaction (HMI)**. While LLM-assisted reviewing is becoming common in clinical research (Scherbakov et al., 2025), its use for the complex, non-standard terminology typical of HMI remains limited. Unlike clinical trials with standardised reporting, HMI design studies on human-automation interaction in HCI often use diverse, qualitative, and non-standardised language. In this context, Asadi et al. show that

standard prompting is often inadequate (Asadi et al., 2025). By iteratively embedding research goals into the prompt, they improved human–AI agreement to 73%, highlighting the need for context-aware instructions rather than simple keyword matching in HMI reviews.

We operationalised the LLM-assisted screening workflow using a Python script. It read 1,178 records with title and abstracts, prompted the model (GPT-5.2) through an API call with the research questions plus explicit eligibility criteria in Section 2.2 and a hierarchical decision rule. Data were gathered between 27 and 28 Jan 2026. The decision rule explicitly states that a paper is excluded if it is outside the eligible interaction scope or if it meets any exclusion criterion (EC1-EC6); otherwise, it is included only when it explicitly addresses inclusion criteria (IC1-IC3), otherwise everything else is excluded. For the output, the model must return a JSON array. The outputs were then exported to a CSV (with the article ID, criteria met, decision, and a brief evidence-based rationale). After screening, the LLM included 54 records and excluded 1,124 records (Figure 2). The agreement between the human reviewer and the LLM at the title and abstract screening was moderate (Cohen's $\kappa = 0.5257$). Of the screened records, **29** were selected by both reviewers, whereas **48** received discrepant decisions: **23** were included only by the human reviewer and **25** only by the LLM. To minimise erroneous exclusions, the union of all potentially eligible **77** records, comprising overlapping and conflicting selections, was advanced to full-text screening for adjudication.

Importantly, we emphasise that the LLM was used to support double screening rather than replace human judgement, consistent with the PRISMA 2020 guidance to transparently report the role of automation tools in study selection.

*2.3.3 Conflict resolution: human and LLM full-text screening.* To reduce the likelihood of mistakenly excluding studies at the title and abstract screening stage, all records for which the first author (human reviewer) and GPT-5.2 disagreed were retained and moved forward to full-text assessment. As a result, the combined set of **77** peer-reviewed articles deemed potentially eligible by at least one of the two reviewers proceeded to full-text screening (Figure 2). Full texts were then evaluated independently using the same inclusion criteria, resulting in substantial agreement between reviewers (Cohen's $\kappa = 0.7996$).

Final decisions on study inclusion were made by the human reviewer, who explicitly analysed the model's rationales and re-read the full texts to substantiate any disagreements. In total, 53 papers were synthesized. Of the **77** articles screened, **47** were selected by both reviewers, while **6** were chosen by the human reviewer but rejected by the LLM reviewer (Chai et al., 2025, Forster et al., 2019, Nobari and Bertram, 2025, Novakazi et al., 2022, Pipkorn et al., 2021, Schwindt-Drews et al., 2024). These papers predominantly focused on trust and behaviour-related outcomes that implicitly diagnose users' mental model about automation (e.g., behavioural errors consistent with misunderstanding or ambiguous system interpretation), despite not explicitly labelling the construct as an inaccurate mental model. We retained these studies because they operationalise relevant constructs and provide conceptual links to user understanding, thereby strengthening the completeness of the KPI inventory we aim to build.

## 2.4 Data extraction process

Data were extracted from the 53 included articles using a structured form developed from the research questions. We proceeded with the extraction at two levels. Firstly, at the factual level, we recorded, as reported, study type, automation level, system feature, experiment setting and driving scenario, sample characteristics, measurement instrument, and operationalised KPIs. The KPI labels were kept close to the operationalisations described by the original authors. The extraction data was gathered between November 2025 and May 2026 and subsequently consolidated into a .csv file.

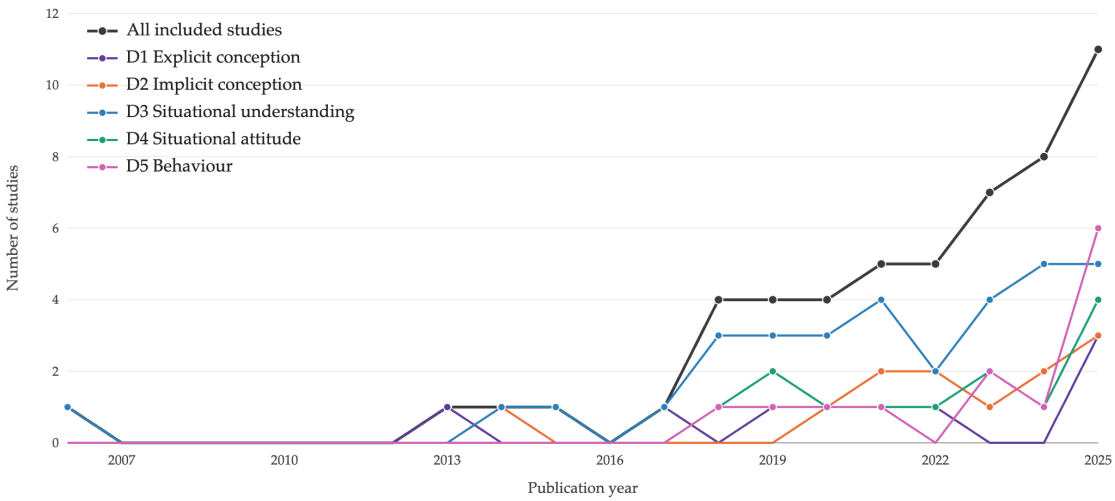


Fig. 3. Full-text reviewed literature to interpret and measure IMM from 2000 to 2025: five dimensions

Secondly, at the interpretive stage, connecting each measure to a KPI, construct, and dimension (D1–D5) represents framework-based judgments by the review team. The complete codebook is available in the supplementary material.

## 3 Results

### 3.1 Study characteristics and framework distribution

Figure 3 traces the temporal distribution of research across the five IMM dimensions over the 2000-2025 review window. Of the 53 included studies, 38 were empirical (71.7%) and 15 were theoretical (28.3%). Together, these studies covered all five dimensions in SAE Levels 1-4. Research activity remained sparse before 2018 but increased markedly thereafter, reaching its highest annual volume in 2025, with 11 publications.

Figure 4 presents the cross-dimensional links among the 53 studies at the level of constructs and KPIs. Each shaded cluster is one dimension; within a cluster, larger nodes are constructs and smaller nodes the KPIs operationalising them, with the number in parentheses giving how many studies used each. Because the coding was multi-label, one article could contribute to several dimensions; consequently, the dimension percentages are non-exclusive and are not expected to sum to 100%. Among the constructs, *mode confusion* (D3) is the most frequently operationalised, featuring in 10 studies, followed by *misuse* (D5, 8 studies), *over-trust* (D4, 5 studies), and *impoverished situation awareness* (D3, 4 studies), *satisfaction*(D4, 4 studies), *under-trust* (D4, 3 studies), and *abuse* in only one. At the KPI level, the pattern is both flatter and sparser: *knowledge accuracy* (D1) is the most common, used in 7 studies, whereas most other KPIs are employed in only one or two. Many studies also span more than one cluster, drawing on cognitive, attitudinal, and behavioural indicators in combination. The framework groups these as complementary evidence.

### 3.2 D1 Explicit conception

Explicit conception externalises the user's internal representation (Norman, 1983a) into declarative knowledge (Anderson, 2013, Rouse and Morris, 1986) that can be checked against an objective ground truth of the system's design specification. Studies in the context of automated driving assess this using scored instruments, where each item has a

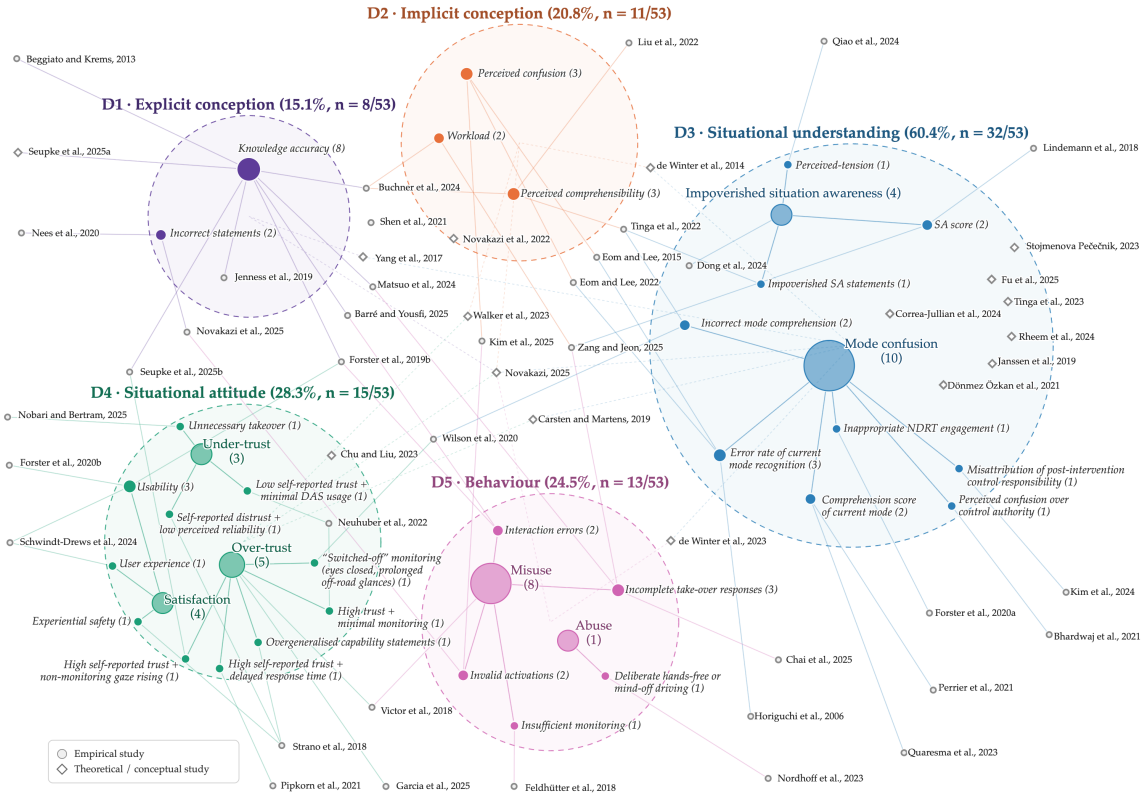


Fig. 4. Hierarchical network of IMM assessment: dimensions (D1–D5), constructs, KPIs and the studies operationalising them.

clearly identifiable correct response tied to particular ADAS features (Beggiato and Krems, 2013, Seupke et al., 2025a). The corresponding KPIs are shown in Table 2, and are applied both objectively and subjectively across studies (N = 10).

On one hand, most studies (N = 7) score responses objectively: items on structured questionnaires assessing *knowledge accuracy* are tied to specific system features and marked against the design specification, even though the concrete implementations vary considerably. Beggiato et al. proposed this paradigm by depicting the mental model of a Level 1 ACC system as a scorable knowledge profile, categorized as correct, incomplete, or incorrect (Beggiato and Krems, 2013). The mental model is assessed using a 35-item instrument, which is summarized into a composite score. Later work has introduced instruments tailored to specific aims: Forster et al. examined four items of operational knowledge with condition-dependent scoring that differentiated L2 from L3 systems (Forster et al., 2019); Buchner proposed a correctness continuum (−3 to +3 plus a "don't know" option) (Buchner et al., 2024); Barré et al. combined a true/false test, a fill-in-the-blank (cloze) task and a diagram-drawing task into a 20-point composite index for a parking assistant (Barré and Yousfi, 2025). Seupke et al. employed a 28-item questionnaire that administers knowledge testing across L2 to L4 (Seupke et al., 2025a). Multiple studies have highlighted the development of mental models over time through longitudinal evaluations (Beggiato and Krems, 2013, Beggiato et al., 2015, Seupke et al., 2025b).

Table 2. D1 Explicit conception: key performance indicators (KPI) and measurements indicating IMM.

| KPIs | Measurements | Study setting | Scenarios | Articles |
|---|---|---|---|---|
| Knowledge accuracy | Questionnaire | Driving simulator | L1 ACC | (Beggiato and Krems, 2013) |
| | Questionnaire | On-road | L2, L3; mode transitions | (Buchner et al., 2024, Forster et al., 2019, Matsuo et al., 2024, Seupke et al., 2025a,b) |
| | Questionnaire composite tasks | Test track | Parking assistant | (Barré and Yousfi, 2025) |
| Incorrect statements | Semi-structured interviews; think-aloud | Qualitative interview study | L1, L2, L3; mode transitions; system-initiated take-over requests | (Jenness et al., 2019, Nees et al., 2020, Novakazi et al., 2025) |

Table 3. D2 Implicit conception: key performance indicators (KPI) and measurements indicating IMM.

| KPIs | Measurements | Study setting | Scenarios | Articles |
|---|---|---|---|---|
| Perceived comprehensibility | Questionnaire, single-item Likert | On-road study | L2 | (Buchner et al., 2024) |
| | Questionnaire, single-item Likert (confidence rating); response time | Online survey | L2 | (Liu et al., 2022) |
| | Questionnaire, multi-item Likert | Online simulated drives | L2, L3, L4; mode transitions | (Tinga et al., 2022) |
| Perceived confusion | Questionnaire, single-item Likert | Driving simulator | L1, L2, L3; mode transitions | (Eom and Lee, 2015, 2022) |
| | Observed or self-reported confusion | Post-hoc interview; observation | L2, L3 | (Kim et al., 2025) |
| Workload | Questionnaire, NASA-TLX | On-road study | L2 | (Buchner et al., 2024) |
| | | Driving simulator | L3 | (Zang and Jeon, 2025) |

On the other hand, explicit conception is accessed in a subjective manner and labeled as "*incorrect statements*," which are derived from open-ended, semi-structured elicitation approaches (Table 2) (N = 3). In these scenarios, participants expressed their own understandings of the system's limitations and its overall interaction logic in relation to the system's HMI (Novakazi et al., 2025), which researchers noted did not align with the ground truth (Jenness et al., 2019). Additionally, researchers highlight several potential origins of these incorrect beliefs: flawed interactions based on past driving experience (Novakazi et al., 2025); exaggerated media portrayals that foster excessive trust in system capabilities (Carsten and Martens, 2019); and misleading terminology, marketing, and social discourse that establish preconceived expectations which carry over into subsequent interaction (Nees et al., 2020, Novakazi et al., 2025).

### 3.3 D2 Implicit conception

Implicit conception refers to the driver's own reported judgement of whether the interaction feels understandable, controllable, and manageable. Its diagnostic power stems from the fact that it is not tied to the objective accuracy of the driver's knowledge, since these measures capture confidence rather than correctness: a driver may claim to fully understand what is happening even when their underlying mental model is clearly misaligned. We found three corresponding KPIs in the literature (N = 7) in Table 3. (i) *perceived comprehensibility* (N = 3) asks drivers how well they believe they grasp the system. Buchner et al., collected two single-item agreement ratings after each L2 drive—one targeting understanding ("I understood the system functionality") and one targeting self-perceived competence ("I have the feeling that I can use the system competently") (Buchner et al., 2024) . Liu et al., operationalised *perceived comprehensibility* as *functional transparency* in an online L2 survey: for each static HMI state, drivers rated how confident they were in that judgement, and how long they took to reach it (Liu et al., 2022). Tinga et al., instead applied a multi-item scale across L2-L4, in which drivers rated how well they understood when they needed to focus on the driving task (Tinga et al., 2022). These items were probed around event-driven mode changes (urban/rural/highway transitions, roadworks, missing lane markings, and fog), anchoring the judgement to concrete situational demands. (ii) *Perceived confusion* (N = 3) captures the driver's own acknowledgement that system state and expectation can no longer be reconciled. Eom et al., used a single post-hoc Likert item to assess how confused drivers were about the system's current mode ("I experienced mode confusion"), administered in scenarios with mode transitions across L1–L3 (Eom and Lee, 2015, 2022). (iii) *Workload* (N = 2) is an indirect index, on the premise that the effort of managing an interaction reflects how well the driver's model supports it. The two empirical studies used NASA-TLX (Hart and Staveland, 1988) which aggregates mental, physical and temporal demand, performance, effort and frustration. Buchner et al., tracked workload alongside objective knowledge accuracy (Table 2) across rising situational complexity (rural to urban), treating it as a companion marker of a developing mental model; notably, however, workload and mental model were analysed in parallel rather than formally linked (Buchner et al., 2024). Zang et al., instead treated workload as a consequence of interface design, indicating that a moderate level of L3 agent's transparency reduced workload and enhanced situation awareness (Zang and Jeon, 2025). Taken together, *workload* functions as a double-edged indicator of IMM: both the *overload* of trying to reconcile the mismatched expectation and the *underload* that comes from disengaging from it can signal an inadequately supported mental model (de Winter et al., 2014).

### 3.4 D3 Situational understanding

*3.4.1 Impoverished situation awareness.* When understood as an observable cue, this construct suggests that the driver's mental model is no longer synchronized with the system state (de Winter et al., 2014) and changing circumstances (de Winter et al., 2023, Fu et al., 2025, Tinga et al., 2023). Impoverished situation awareness consistently appears as a distinct construct in the reviewed literature (N = 13), with the majority of these contributions being conceptual or review articles (de Winter et al., 2014, Pečečnik, 2023).

We found three corresponding KPIs in Table 4 across empirical literature (N = 4). (i) *SA score*, derived from SAGAT freeze probes, is measured by disrupting the drive and evaluating drivers' responses about the situation against ground truth across Endsley's three levels—perception, comprehension, and projection (Endsley, 1995, 2015). This method has been used at predefined critical moments in L4 urban scenarios (Lindemann et al., 2018) as well as during take-over situations in L3 (Zang and Jeon, 2025). (ii) *Impoverished situation awareness statements* are captured through concurrent think-aloud across L2–L4 traffic and mode transitions, coding utterances that reveal incorrect awareness through

Table 4. D3 Situational understanding (impoverished situation awareness): key performance indicators (KPI) and measurements indicating IMM.

| KPIs | Measurements | Study setting | Scenarios | Articles |
|---|---|---|---|---|
| SA score | Questionnaire SAGAT, freeze-probe | Driving simulator | L3, L4; | (Lindemann et al., 2018, Zang and Jeon, 2025) |
| Impoverished situation awareness statements | Concurrent think-aloud | Online simulated drives | L2, L3, L4; mode transitions | (Tinga et al., 2022) |
| Perceived-tension | Questionnaire, single-item | Driving simulator | L2; warning; take-over | (Qiao et al., 2024) |

Table 5. D3 Situational understanding (mode confusion): key performance indicators (KPI) and measurements indicating IMM.

| KPIs | Measurements | Study setting | Scenarios | Articles |
|---|---|---|---|---|
| Error rate of current mode recognition | Self-reported mode answers; freeze-probe | Driving simulator | L1 ACC | (Eom and Lee, 2015, Horiguchi et al., 2006) |
| | Self-reported mode answers; freeze-probe | Driving simulator | L1, L2, L3; ACC, LKA, system initiated lane change (SI-LC); | (Eom and Lee, 2022) |
| Comprehension score of current mode | Researcher rating open-ended descriptions | Online survey | L1, L2, L3; ACC, LKA | (Perrier et al., 2021) (Quaresma et al., 2023) |
| Perceived confusion over control authority | Questionnaire, single-item Likert | Driving simulator | L2, L3 | (Bhardwaj et al., 2021) |
| Misattribution of post-intervention control responsibility | Questionnaire | Online survey | L2, L3; | (Kim et al., 2024) |
| Inappropriate NDRT engagement | Count of finger inputs on a secondary task | Driving simulator | L2, L3 | (Forster et al., 2020a) |
| Incorrect mode comprehension | Observation; post-hoc interview | On-road study | L2 | (Wilson et al., 2020) |
| | Concurrent think-aloud | Online simulated drives | L2,L3,L4; event-driven mode changes | (Tinga et al., 2022) |

verbal coding (Tinga et al., 2022). (iii) *Perceived-tension* is employed by Qiao et al. as a post-hoc, single-item measure to quantify the deviation between drivers' perceived tension and the actual level of tension in L2 warning and take-over scenarios. (Qiao et al., 2024).

*3.4.2 Mode confusion.* Mode confusion recurred as a distinct construct across the reviewed literature (N = 19). Here *mode* is not merely the nominal level of driving automation but the current configuration of the driver-system interaction:

which functions are active, and how control authority and monitoring responsibility are allocated between driver and system (Janssen et al., 2019, Kim et al., 2025, Kurpiers et al., 2020, Parasuraman et al., 2000). A mismatch on either facet counts as mode confusion (Carsten and Martens, 2019, Sarter and Woods, 1995). Prior work has linked this mismatch to complex mode transition logic (Carsten and Martens, 2019, de Winter et al., 2023), has identified IMM as a primary source of the problem (Eom and Lee, 2022), or has introduced formal approaches to estimate when drivers' awareness and system state become misaligned (Rheem et al., 2024).

Among the remaining empirical evidence (N = 10), we found six corresponding KPIs (Table 5), which differ in whether the mismatch is probed, rated, or inferred. IMM, therefore, registers as a gap between the assumed and the actual system feature, particularly in terms of identity, how responsibility is assigned, and what permissions are in place. The error rate of current-mode recognition probes it directly: the drive is frozen, at random (Eom and Lee, 2015, Horiguchi et al., 2006) or after scripted events (Eom and Lee, 2022), and the self-reported mode is scored against the actual one (Eom and Lee, 2022). Comprehension scores instead of current mode rate participants' open-ended descriptions of modes and HMI symbols against researcher criteria (Perrier et al., 2021, Quaresma et al., 2023). Two questionnaire measures mode confusion through testing if participants understand the control authority under specific modes: a single-item rating of perceived confusion over control authority (Bhardwaj et al., 2021), and a large-scale scenario survey (N = 838) that benchmarks the share of respondents attributing speed, distance and steering control to "driver" versus "car" after each intervention against commercial transition logic (Kim et al., 2024). The remaining KPIs infer confusion from behaviour or verbal statements. Forster et al. used the count of finger inputs on a secondary task indexes drivers' mode awareness (Forster et al., 2020a). In on-road observation, post-drive interviews (Wilson et al., 2020) and concurrent think-aloud coding (Tinga et al., 2022) captured mode confusion, such as drivers assuming automation remained engaged after an unintended disengagement.

### 3.5 D4 Situational attitude

This dimension captures drivers' momentary attitudinal states during the interaction with the system (Gold et al., 2015, Lee and See, 2004) and recurred across the reviewed literature (N = 15) through three constructs: *under-trust*, *over-trust*, and *satisfaction*. Instead of functioning as a straightforward indicator, attitude is used to infer the state of the mental model (Hoff and Bashir, 2015, Lee and See, 2004). Its diagnostic utility stems from calibration (Walker et al., 2023)—whether the attitude aligns with the system's actual capabilities. We found three constructs (Table 6) that each capture a distinct manner in which this calibration can break down. Across the empirical studies (N = 9), eleven matching KPIs were identified, all of which involved combining a subjective rating with an objective channel that could serve as a reference for evaluating IMM.

*3.5.1 Under-trust.* This construct denotes reliance that is inappropriately low relative to the system's demonstrated capability, and thereby reveals an IMM that underrates automation competence or exaggerates situational risk. Its KPIs take two forms in empirical studies (N = 3): reduced reliance on a capable system, and over-intervention in benign situations (Table 6). Neuhuber et al. operationalised it by low trust ratings coinciding with minimal system usage, identifying that under-trust appeared when participants declined the system when it was demonstrably capable (Neuhuber et al., 2022). Nobari et al. additionally performed real-time integrated measurements, recording from the simulator all instances of braking, steering, or other control inputs made during automated L2 car-following that were objectively non-critical, and pairing these with elevated skin-conductance responses—i.e., increases in the skin's electrical conductance caused by involuntary sweating under arousal and captured via electrodermal activity (EDA) (Nobari and Bertram, 2025).

Table 6. D4 Situational attitude: key performance indicators (KPI) and measurements indicating IMM.

| KPIs | Measurements | Study setting | Scenarios | Articles |
|---|---|---|---|---|
| **Under-trust** | | | | |
| Low self-reported trust + minimal DAS usage | Questionnaire; behaviour log | On-road study | L2 | (Neuhuber et al., 2022) |
| Unnecessary takeover | Takeover logs (binary); SCR amplitude (EDA) | Driving simulator | L2; non-critical events | (Nobari and Bertram, 2025) |
| Self-reported distrust + low perceived reliability | Semi-structured interview | Interview study | L2, L3 | (Strano et al., 2018) |
| **Over-trust** | | | | |
| High self-reported trust + non-monitoring gaze rising | Questionnaire; eye-tracking ROI | On-road | L3 | (Seupke et al., 2025b) |
| High self-reported trust + delayed response time | Questionnaire; behaviour log by timestamps | On-road | L2; conflict events | (Pipkorn et al., 2021) |
| High trust + minimal monitoring | Questionnaire; eye-tracking | On-road study | L2, L3 | (Neuhuber et al., 2022) |
| Overgeneralised capability statements | Semi-structured interview | Interview study | L2, L3 | (Strano et al., 2018) |
| "Switched-off" monitoring (eyes closed, prolonged off-road glances) | Observation; post-hoc interview | On-road study | L2 | (Wilson et al., 2020) |
| **Satisfaction** | | | | |
| Usability | Questionnaire (SUS), multi-item | Driving simulator; on-road study | L2, L3 | (Forster et al., 2019, 2020b, Schwindt-Drews et al., 2024) |
| User experience | Questionnaire (UEQ), multi-item | On-road study | L3 | (Schwindt-Drews et al., 2024) |
| Experiential safety | Semi-structured interview; self-reported emotion delta | Workshop, interview study | L2, L3 | (Strano et al., 2018) |

*3.5.2 Over-trust.* This construct refers to reliance that exceeds system limitations and reflects an overly optimistic or boundary-insensitive mental model (e.g., assuming the system can manage scenarios outside its operational design domain) (Seupke et al., 2025b, Strano et al., 2018). Automation complacency is the attentional consequence of such patterns, where prolonged reliable performance produces declines in vigilance (Chu and Liu, 2023, de Winter et al., 2023, Victor et al., 2018). Its KPIs take three forms in empirical studies (N = 5), escalating from belief to consequence: overgeneralised capability beliefs, withdrawn supervision under high trust, and delayed intervention when the system reaches its limits (Table 6). The first appears in interview statements assuming the system can handle situations beyond its actual envelope (Strano et al., 2018). The second pairs high trust ratings with objectively reduced monitoring: learned trust (LETRAS-G) rose together with non-monitoring gaze over four weeks of on-road L3 use (Seupke et al., 2025b); high ratings co-occurred with minimal monitoring that barely adapted as risk increased (Neuhuber et al., 2022);and, at

the extreme end, observers recorded instances of "switched-off" behaviour—closed eyes, extended glances away from the road—which participants themselves attributed to overconfidence (Wilson et al., 2020). The third form captures the safety-critical consequence: in an on-road cut-out conflict event, high-trust drivers awaited an intervention the system could not deliver and therefore responded later along the video-coded chain of surprise reaction, hands-on-wheel, steering and braking, with the eventual "crashers" showing the longest delays (Pipkorn et al., 2021). Pipkorn et al. attribute the delay to an expectation the system's capability could not meet.

*3.5.3 Satisfaction.* This construct reflects how users subjectively judge the quality of the interaction. Instead of capturing reliance or knowledge directly, it bears on whether the HMI supports a workable mental model. Forster et al. read low usability after mode transitions as a sign that the interface failed to convey how the system works. (Forster et al., 2019, 2020b). The corresponding KPIs represent three types of experience quality: *usability*, *user experience* and *experiential safety* in empirical studies (N = 4). Usability is measured after driving with the ten-item System Usability Scale (SUS) (Brooke and others, 1996), administered after simulator-based L2/L3 mode transitions (Forster et al., 2019, 2020b) and after first contact with a L3 system in real traffic (Schwindt-Drews et al., 2024); the latter study also rated user experience with the User Experience Questionnaire (UEQ) (Laugwitz et al., 2008). Experiential safety was assessed qualitatively: during board-game workshops that simulated L2/L3 scenarios with unexpected incidents, participants rated their emotional states, which were identified as frustration, anxiety, and perceived lack of safety that indicated expectation violation (Strano et al., 2018).

### 3.6 D5 Behaviour

This dimension illustrates observable outcomes of interactions, highlighting how users apply their understanding of automation during real driving tasks. This concept recurs throughout the literature that we reviewed (N = 13) and is divided into two main constructs based on intent: (i) *misuse*, which occurs when operational limits are breached unknowingly, and (ii) *abuse*, where these limits are deliberately violated (Parasuraman and Riley, 1997). Behavioural deviations reveal the moments directly when the IMM of a driver fails to guide actions effectively. We identified five corresponding KPIs across empirical studies (N = 10) as shown in Table 7.

*3.6.1 Misuse.* This construct makes IMM observable at the moment when an inaccurate belief authorizes an action that the system does not endorse. Four KPIs were identified in the reviewed literature (N = 9), spanning successive stages of automation use: invalid activations, insufficient monitoring, incomplete take-over responses, and interaction errors (Table 7). The first captures misuse at the boundary of the operational design domain. During on-road Wizard-of-Oz L2/L3 drives, in-vehicle video recorded drivers' repeated attempts to activate automation where its use was not permitted. Measurements were conducted with think-aloud protocols and post-drive interviews that linked these attempts to specific misconceptions (Kim et al., 2025, Novakazi et al., 2025). The second KPI captures the misuse of the supervisory role. After automation was downgraded from L3 to L2, drivers continued to behave as though the permissions of the previous mode were still applied: they maintained the participation in secondary-tasks, kept their gaze away from the road, responded late or not at all to automation errors, and made more false interventions (Feldhütter et al., 2018). The third captures incomplete or delayed take-over responses via simulator data (de Winter et al., 2023). Chai et al. conducted it in a critical motorway cut-in with a four-second take-over budget, simulator logs registered steering without braking, delayed brake onset, and unstable steering, the latter measured by the standard deviation of steering-wheel angle during the 20s following the request (Chai et al., 2025). The fourth KPI concerns errors in executing system interactions. In simulator studies, these were measured through logged wrong button presses and

Table 7. D5 Behaviour: key performance indicators (KPI) and measurements indicating IMM.

| KPIs | Measurements | Study setting | Scenarios | Articles |
|---|---|---|---|---|
| **Misuse** | | | | |
| Invalid activations | Behaviour logs; concurrent think-aloud; post-drive interviews | On-road study | L2, L3 | (Kim et al., 2025, Novakazi et al., 2025) |
| Insufficient monitoring | Simulator logs; eye-tracking | Driving simulator | L2, L3; mode transition | (Feldhütter et al., 2018) |
| Incomplete take-over responses | Simulator logs | Driving simulator | L2, L3; critical take-over (motorway cut-in scenarios) | (Chai et al., 2025, Zang and Jeon, 2025) |
| | Simulator logs | Driving simulator | L2; take-over when automation fails (curves, fog scenarios) | (Matsuo et al., 2024) |
| Interaction errors | Simulator logs (wrong button presses, timing); Researcher rating interaction | Driving simulator | L2, L3; mode transitions | (Forster et al., 2019, 2020b) |
| | Video-coded error counts | Test track | Parking assistant | (Barré and Yousfi, 2025) |
| **Abuse** | | | | |
| Deliberate hands-free or mind-off driving | Semi-structured interview | Interview study | L2 (Tesla Autopilot/FSD) | (Nordhoff et al., 2023) |

task-completion times, supplemented by researcher ratings of interaction success (Forster et al., 2019, 2020b); on a test track, they were measured as video-coded error counts during repeated use of a parking assistant (Barré and Yousfi, 2025). The decline of these errors with training and repeated experience indicates that they reflect remediable knowledge deficits rather than stable behavioural dispositions (Barré and Yousfi, 2025, Forster et al., 2019).

*3.6.2 Abuse.* This construct differs from misuse in that the operator deliberately breaches operational limits rather than misinterpreting them across studies (N = 1). Nordhoff et al. identified such behaviour via semi-structured interviews with Tesla FSD users (Nordhoff et al., 2023), documenting purposeful hands-free driving in situations where it is not allowed and deliberate "mind-off" driving in which drivers consciously withdraw from supervisory duties. From an IMM standpoint, abuse constitutes a borderline case: it may reflect a correct understanding of system capabilities combined with a deliberate choice to operate beyond safe limits. Due to this ambiguity, behavioural data alone cannot reliably differentiate abuse from misuse without additional evidence regarding the user's underlying knowledge state.

## 4 Discussion

The principal contribution of this review is a framework that organises how IMM is conceptualised and assessed in human-DAS interaction (Figure 5). Overall, the review delineated 7 constructs and 30 KPIs, and linked them to their relevant measurement instruments, automation levels, and traffic conditions. In conclusion, we situate the findings

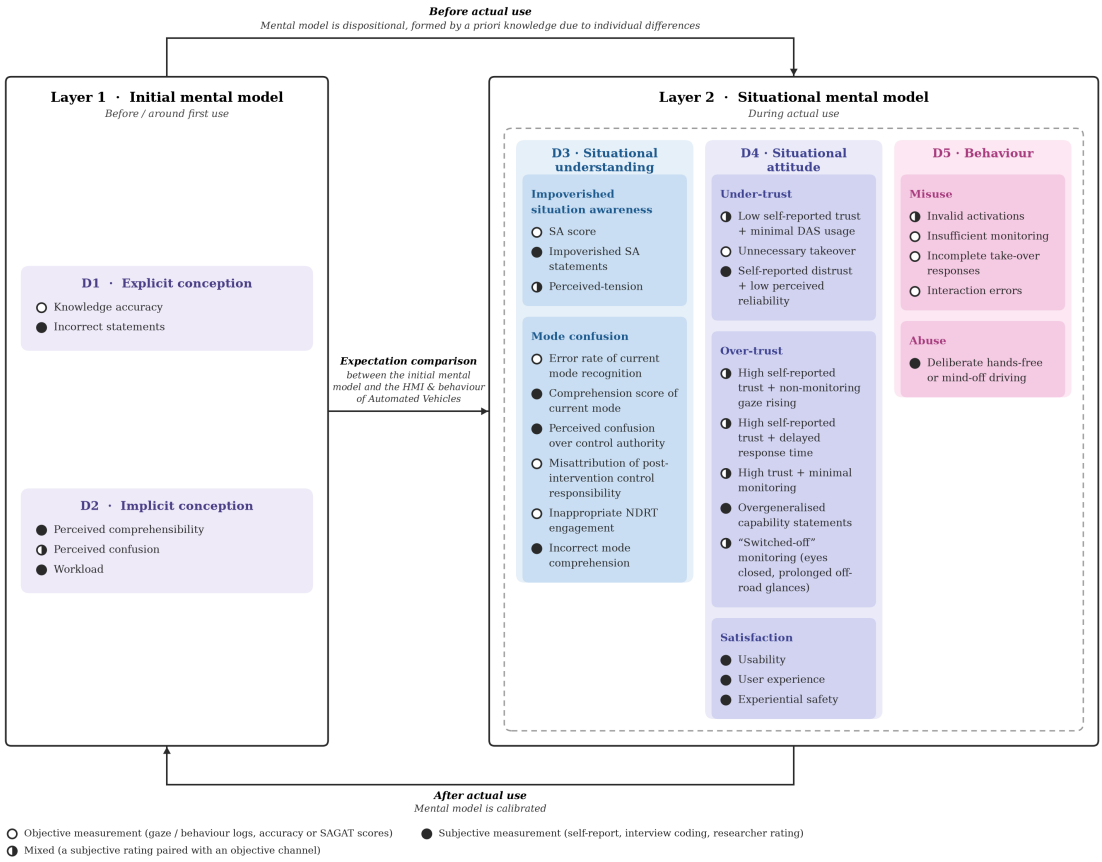


Fig. 5. IMM diagnostic constructs and KPIs derived from the five-dimension framework.

within this framework by explicitly interpreting constructs and classifying KPIs, which can be evaluated relative to a defined system state or capability. These direct and proxy indicators still provide valuable evidence and underpin a defensible IMM diagnosis, supported by corresponding measures of the driver's beliefs and converging evidence that rules out plausible alternative explanations.

### 4.1 RQ1: How is IMM conceptualised?

RQ 1 asked *how is inaccurate mental model (IMM) of driving automation conceptualised in the reviewed literature*, that is, which constructs are used to characterise it and how are they differentiated? The synthesis shows that these constructs describe different locations in a connected process rather than mutually exclusive outcomes. It reveals a key challenge in our reviewing process: the constructs used to interpret IMM in practice partially explain, mediate, or operationalise each other. For instance, within **D3 situational understanding**, we distinguish the constructs using terminology in human factors (de Winter et al., 2023): (i) impoverished situation awareness (de Winter et al., 2014) encompasses an extensive "out-of-the-loop" perception following perceptual–comprehension–projection chain (Endsley, 1995). It

manifests as generic cognitive disengagement due to IMM in the human-DAS interaction (de Winter et al., 2014, 2023, Qiao et al., 2024, Tinga et al., 2022, 2023); (ii) "*mode confusion*" (Carsten and Martens, 2019, Rushby et al., 1999) provides a narrower diagnostic window focused on whether the driver's situational belief about the current system state matches its actual state (Eom and Lee, 2015, 2022, Kim et al., 2025, 2024). However, the taxonomy differentiates constructs by their theoretical locus, that is, the particular cognitive function each is intended to diagnose, rather than by the instruments used to measure them. Accordingly, we argue that researchers employing the assessment inventory should be aware that relying on a single construct–KPI pairing creates a risk of misattribution. For instance, A "*delayed takeover*" response captured as a behavioural KPI in *D5 behaviour* cannot, on its own and without accompanying cognitive or attitudinal evidence from *D3 situational understanding* or *D4 situational attitude*, reliably discriminate between "*mode confusion*", "*impoverished situational awareness*", *over-trust*, or the more general notion of expectation violation that helps to account for all of these. Therefore, we emphasise that this framework functions not only as a classification scheme but also as a decision-support device, indicating when multi-dimensional triangulation is required for a valid diagnosis.

### 4.2 RQ2: How is IMM operationalised and applied?

RQ2 asked *how is IMM of driving automation operationalised and applied in the reviewed literature*, that is, which KPIs and measurement methods are used to make it observable? The inventory contains 30 KPI categories: 9 objective (30.0%), 13 subjective (43.3%), and 8 combining objective and subjective evidence (26.7%). Among the 38 empirical studies with a coded analysis approach, 26 were quantitative (68.4%), 9 used mixed methods (23.7%), and 3 were qualitative (7.9%). These results reveal a measurement-validity gap. Subjective measures reveal what drivers believe or how they explain an event, whereas behaviour logs, gaze, physiological signals, and ground-truth-scored questionnaires show what they detect or do. Either channel can be misleading alone, and divergence between them may itself be diagnostic (Neuhuber et al., 2022).

Our synthesis brings together three knowledge gaps. Firstly, individual KPIs are operationalised internally and offer only limited comparability across different studies. The construct of *mode confusion* in *D3 situational understanding* illustrates this most acutely. The reviewed studies make use of fairly well-targeted measures, including freeze-trial mode recognition tasks (Eom and Lee, 2015, Horiguchi et al., 2006), event-triggered self-reports of current mode status (Eom and Lee, 2022), researcher-assigned comprehension scores (Perrier et al., 2021, Quaresma et al., 2023), and questions about responsibility allocation (Kim et al., 2024). However, each of these approaches is closely linked to the particular system's mode architecture, which in turn is shaped by its branding, HMI design, and the specific experimental configurations used. For instance, a mode-recognition error rate from a random-freeze protocol in an L1 ACC simulator study (Eom and Lee, 2015) is not directly comparable to one derived from event-triggered freezes across L1–L3 multi-mode transitions (Eom and Lee, 2022).

Secondly, KPIs were operationalised in imbalanced scenarios. Of the 38 empirical studies, 14 were motorway-based (36.8%), 7 examined mixed urban and motorway traffic (18.4%), and only 4 focused on urban traffic (10.5%); the remaining 12 (34.2%) formed a heterogeneous "Other" category including online, static-interface, closed-track, parking, or otherwise non-comparable settings. Automation-level coverage was also concentrated: the non-exclusive counts were 29 studies for L2 (76.3%), 19 for L3 (50.0%), 9 for L1 (23.7%), and 4 for L4 (10.5%). This imbalance is not merely a sampling limitation but a diagnostic one: different scenario types selectively activate different dimensions of the framework. Motorway-like scenarios involve sustained, steady-state automation with infrequent mode transitions, making them effective at exposing knowledge of activation conditions and takeover readiness (Chai et al., 2025, Forster et al., 2019, Novakazi et al., 2025). Urban traffic, by contrast, generates continuous ambiguity, raising denser interaction with vulnerable

road users, socially negotiated right-of-way, frequent intent changes, and automation behaviours such as slowing, yielding, or hesitating. That shifts the diagnostic challenge from mode recognition toward situational comprehension of not only *what* the system is doing but *why*. Current KPIs, which are largely built around binary state-detection tasks or single-event response measures, appear too coarse-grained to reflect this level of nuance. This shortcoming is particularly problematic for L3 systems or above, which are being deployed with increasing frequency in urban operational design domains but are still assessed mainly under motorway conditions. None of the studies in the reviewed corpus explicitly conceptualises scenario selectivity, indicating that scenario selection is still treated as a matter of convenience in experimental design rather than as a diagnostic choice aligned with the construct being examined.

A further gap concerns temporality. Most of the available evidence comes from single-session simulator tasks or isolated drives; however, a subset, 26.3% (N=10) of 38 experimental studies, indicates that the precision of the mental model does not remain stable once formed initially. Seupke et al. found that dynamically evolving learned trust and mental model development unfold together during repeated real-world use of L3 automation (Seupke et al., 2025b), and Beggiato and Krems observed that, with repeated exposure to ACC, drivers' mental model profiles gradually converge toward greater accuracy (Beggiato and Krems, 2013). Together, these results suggest that observable behavioural competence can conceal persistent misunderstanding and that misconceptions may endure or reappear in novel contexts even after models appear calibrated. Consequently, future IMM research should address not only whether a driver's mental model is inaccurate at a particular point in time, but also how that inaccuracy emerges, stabilises, or reoccurs over the course of the interaction history.

### 4.3 Using LLM in systematic review

Incorporating an LLM as a supportive screener into a PRISMA-compliant study selection workflow can enhance the efficiency of human reviewers, although it still exhibits several recurring limitations that should be refined in future research. Its errors were nonetheless systematic. Firstly, the LLM tended to rely heavily on superficial cues under limited context of screening titles and abstracts, in contrast to a human expert with established knowledge of human factors in automated driving. For example, it disproportionately favored work labeled as "human–machine" or "shared control," and at times included engineering-oriented papers that mentioned these terms in the abstract but did not present any human-factors evidence. The model also showed weak handling of absence-of-evidence, at times inferring plausible but unsubstantiated details. This issue was particularly evident when abstracts used vague terms (e.g., "awareness" to indicate vigilance rather than understanding), a distinction the deployed model cannot detect within the context of our review criteria. Moreover, domain-specific conventions—for instance, interpreting participants' grasp of system symbols as evidence of their underlying mental models without explicitly invoking terms like "mode confusion", were not reliably captured by the model. This inconsistency indicates that the effectiveness of screening still depended strongly on the exact wording of prompts and the disciplinary context in which they were used, particularly in niche fields such as HMI where terminology lacks standardisation.

## 5 Limitations and future work

Several limitations should be acknowledged. First, study selection and limitation of LLM-screening performance was observed on GPT-5.2, OpenAI's strongest available model at the time (OpenAI, 2025). Later releases, including GPT-5.6 and Claude Fable 5, report improved reasoning and long-context performance (Anthropic, 2026, OpenAI, 2026), suggesting that some failure modes may now be less frequent. This possibility nevertheless requires task-specific validation because screening performance varies across models, prompts, and datasets in future work (Dennstädt

et al., 2024, Hilkenmeier et al., 2026). Second, the five-dimension framework and the construct–KPI inventory were derived synthetically from the reviewed literature rather than empirically validated through independent data collection; the taxonomy reflects how the field currently operationalises IMM, not a ground-truth cognitive architecture. Third, construct assignment required interpretive judgement, particularly for studies that did not explicitly frame their measures as mental-model diagnostics, meaning that alternative coding decisions could yield a somewhat different inventory. Finally, the review scope was bounded to in-vehicle human–automation interaction at SAE L1–L4, excluding exterior communication (eHMI), remote operation, and passenger-only contexts, which limits generalisability to the broader automated mobility ecosystem.

## Supplementary material

The screening record set, the paired human and LLM screening decisions csv files, the complete codebook, and the relevant coding scheme are available for peer review in an anonymised link (https://www.dropbox.com/scl/fo/0r709awp695257vv1pqxv/AGiCYHhnxRTX1-XlD32ENFw?rlkey=b9wt27mm1p7ucx3sx90a4blc6&st=xd5j5rfu&dl=0) and are archived under a persistent DOI supplied to the editorial office. Both the dataset and a maintained version of the code will be made publicly available upon acceptance.

## Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used a large language model (GPT-5.2) as an independent second reviewer during title, abstract and full-text screening; this use is part of the study method and is described in full in Section 2. In addition, the authors used a large language model to check grammar and to improve the readability of the manuscript text. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

## References

John R Anderson. 2013. *The architecture of cognition.* Psychology Press.

Anthropic. 2026. *Claude fable 5 and claude mythos 5 system card.* Technical Report. Anthropic. https://www-cdn.anthropic.com/2f9323abbcc4abe219577539efe19a623c9ca2bd/Claude%20Fable%205%20%26%20Claude%20Mythos%205%20System%20Card.pdf

Amir Reza Asadi, Yuchong Zhang, and Hazem Said. 2025. What do personas say about privacy & security: a systematic literature review through human-AI collaboration. *Journal of Ambient Intelligence and Humanized Computing* 16, 11-12 (Dec. 2025), 1175–1190. doi:10.1007/s12652-025-05007-w

Petra Badke-Schaub, Andre Neumann, Kristina Lauche, and Susan Mohammed. 2007. Mental models in design teams: a valid approach to performance in design collaboration? *CoDesign* 3, 1 (March 2007), 5–20. doi:10.1080/15710880601170768 _eprint: https://doi.org/10.1080/15710880601170768.

Victoria A. Banks, Alexander Eriksson, Jim O'Donoghue, and Neville A. Stanton. 2018. Is partially automated driving a bad idea? Observations from an on-road study. *Applied Ergonomics* 68 (April 2018), 138–145. doi:10.1016/j.apergo.2017.11.010

Jessy Barré and Elsa Yousfi. 2025. Training older drivers to use automatic parking assistance: Error analysis and mental model evolution. *Journal of Transport & Health* 44 (Oct. 2025), 102098. doi:10.1016/j.jth.2025.102098

Matthias Beggiato and Josef F. Krems. 2013. The evolution of mental model, trust and acceptance of adaptive cruise control in relation to initial information. *Transportation Research Part F: Traffic Psychology and Behaviour* 18 (May 2013), 47–57. doi:10.1016/j.trf.2012.12.006

Matthias Beggiato, Marta Pereira, Tibor Petzoldt, and Josef F. Krems. 2015. Learning and development of trust, acceptance and the mental model of ACC. A longitudinal on-road study. *Transportation Research Part F: Traffic Psychology and Behaviour* 35 (Nov. 2015), 75–84. doi:10.1016/j.trf.2015.10.005

Akshay Bhardwaj, Yidu Lu, Selina Pan, Nadine B. Sarter, and R. Brent Gillespie. 2021. Comparing Coupled and Decoupled Steering Interface Designs for Emergency Obstacle Evasion. *IEEE Access* 9 (2021), 116857–116868. doi:10.1109/ACCESS.2021.3106225

BMW Group. 2023. Level 3 highly automated driving available in the new BMW 7 Series from next spring. https://www.press.bmwgroup.com/global/article/detail/T0438214EN/level-3-highly-automated-driving-available-in-the-new-bmw-7-series-from-next-spring?language=en tex.howpublished: Press release.

Anika Boelhouwer, Arie P. van den Beukel, Mascha C. van der Voort, and Marieke H. Martens. 2019. Should I take over? Does system knowledge help drivers in making take-over decisions while driving a partially automated car? *Transportation Research Part F: Traffic Psychology and Behaviour* 60 (Jan. 2019), 669–684. doi:10.1016/j.trf.2018.11.016

Fabiano Tonaco Borges, Gabriela do Manco Machado, Maíra Araújo de Santana, Karla Amorim Sancho, Giovanny Vinícius Araújo de França, Wellington Pinheiro dos Santos, and Carlos Eduardo Gomes Siqueira. 2026. LLM-Assisted Scoping Review of Artificial Intelligence in Brazilian Public Health: Lessons from Transfer and Federated Learning for Resource-Constrained Settings. *International Journal of Environmental Research and Public Health* 23, 1 (Jan. 2026), 81. doi:10.3390/ijerph23010081

John Brooke and others. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

Claudia Buchner, Chantal Himmels, Jan Schmitz, Tanja Stoll, and Martin Baumann. 2024. Exploring Urban Challenges: Understanding Advanced Driver Assistance Systems in Different Situational Contexts. In *Proceedings of the 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, Stanford CA USA, 1–12. doi:10.1145/3640792.3675709

Oliver Carsten and Marieke H. Martens. 2019. How can humans understand their automated cars? HMI principles, problems and solutions. *Cognition, Technology & Work* 21, 1 (Feb. 2019), 3–20. doi:10.1007/s10111-018-0484-0

Chen Chai, Shixuan Weng, Rui Feng, Jixiang Wang, Yiik Diew Wong, and Yanbo Wang. 2025. A behavioral inoculation method to improve critical takeover performance of novice drivers in conditionally automated driving. *Traffic Injury Prevention* 26, 6 (Aug. 2025), 642–652. doi:10.1080/15389588.2025.2451699

Sully F. Chen, Anton Alyakin, Andreas Seas, Eunice Yang, Joanne J. Choi, Jin Vivian Lee, Amelia L. Chen, Pranav I. Warman, Rochelle T. Bitolas, Robert J. Steele, Daniel A. Alber, and Eric K. Oermann. 2026. LLM-assisted systematic review of large language models in clinical medicine. *Nature Medicine* 32, 3 (March 2026), 1152–1159. doi:10.1038/s41591-026-04229-5

Yueying Chu and Peng Liu. 2023. Automation complacency on the road. *Ergonomics* 66, 11 (Nov. 2023), 1730–1749. doi:10.1080/00140139.2023.2210793 _eprint: https://doi.org/10.1080/00140139.2023.2210793.

Allan Collins and Dedre Gentner. 1987. How people construct mental models. *Cultural models in language and thought* 243, 1987 (1987), 243–265. doi:10.1017/CBO9780511607660.011

European Transport Safety Council. 2026. Letter: Questions on RDW's Approval of Tesla FSD Supervised. (2026). https://etsc.eu/letter-questions-on-rdws-approval-of-tesla-fsd-supervised/

Kenneth James Williams Craik. 1967. *The nature of explanation*. Vol. 445. CUP Archive.

Joost C.F. de Winter, Riender Happee, Marieke H. Martens, and Neville A. Stanton. 2014. Effects of adaptive cruise control and highly automated driving on workload and situation awareness: A review of the empirical evidence. *Transportation Research Part F: Traffic Psychology and Behaviour* 27 (Nov. 2014), 196–217. doi:10.1016/j.trf.2014.06.016

Joost C.F. de Winter, Sebastiaan M. Petermeijer, and David A. Abbink. 2023. Shared control versus traded control in driving: a debate around automation pitfalls. *Ergonomics* 66, 10 (Oct. 2023), 1494–1520. doi:10.1080/00140139.2022.2153175

Fabio Dennstädt, Johannes Zink, Paul Martin Putora, Janna Hastings, and Nikola Cihoric. 2024. Title and abstract screening for literature reviews using large language models: an exploratory study in the biomedical domain. *Systematic Reviews* 13, 1 (June 2024), 158. doi:10.1186/s13643-024-02575-4

Francis T. Durso, Katherine A. Rawson, and Sara Girotto. 2007. Comprehension and Situation Awareness. In *Handbook of Applied Cognition*. John Wiley & Sons, Ltd, 163–193. doi:10.1002/9780470713181.ch7 Section: 7 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470713181.ch7.

Mica R. Endsley. 1995. Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors* 37, 1 (March 1995), 32–64. doi:10.1518/001872095779049543

Mica R. Endsley. 2015. Situation Awareness Misconceptions and Misunderstandings. *Journal of Cognitive Engineering and Decision Making* 9, 1 (March 2015), 4–32. doi:10.1177/1555343415572631

Hwisoo Eom and Sang Hun Lee. 2015. Human-Automation Interaction Design for Adaptive Cruise Control Systems of Ground Vehicles. *Sensors* 15, 6 (June 2015), 13916–13944. doi:10.3390/s150613916

Hwisoo Eom and Sang Hun Lee. 2022. Mode confusion of human–machine interfaces for automated vehicles. *Journal of Computational Design and Engineering* 9, 5 (Oct. 2022), 1995–2009. doi:10.1093/jcde/qwac088

Anna Feldhütter, Christoph Segler, and Klaus Bengler. 2018. Does Shifting Between Conditionally and Partially Automated Driving Lead to a Loss of Mode Awareness? In *Advances in Human Aspects of Transportation*, Neville A Stanton (Ed.). Vol. 597. Springer International Publishing, Cham, 730–741. doi:10.1007/978-3-319-60441-1_70 Series Title: Advances in Intelligent Systems and Computing.

Yannick Forster, Viktoria Geisel, Sebastian Hergeth, Frederik Naujoks, and Andreas Keinath. 2020a. Engagement in Non-Driving Related Tasks as a Non-Intrusive Measure for Mode Awareness: A Simulator Study. *Information* 11, 5 (April 2020), 239. doi:10.3390/info11050239

Yannick Forster, Sebastian Hergeth, Frederik Naujoks, Matthias Beggiato, Josef F. Krems, and Andreas Keinath. 2019. Learning to use automation: Behavioral changes in interaction with automated driving systems. *Transportation Research Part F: Traffic Psychology and Behaviour* 62 (April 2019), 599–614. doi:10.1016/j.trf.2019.02.013

Yannick Forster, Sebastian Hergeth, Frederik Naujoks, Josef F. Krems, and Andreas Keinath. 2020b. What and how to tell beforehand: The effect of user education on understanding, interaction and satisfaction with driving automation. *Transportation Research Part F: Traffic Psychology and Behaviour* 68 (Jan. 2020), 316–335. doi:10.1016/j.trf.2019.11.017

Qianwen Fu, Lijun Zhang, Yiqian Xu, and Fang You. 2025. The Review of Human–Machine Collaborative Intelligent Interaction With Driver Cognition in the Loop. *Systems Research and Behavioral Science* 42, 4 (2025), 954–977. doi:10.1002/sres.3141 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sres.3141.

Gerald Gartlehner, Lisa Affengruber, Viktoria Titscher, Anna Noel-Storr, Gordon Dooley, Nicolas Ballarini, and Franz König. 2020. Single-reviewer abstract screening missed 13 percent of relevant studies: a crowd-based, randomized controlled trial. *Journal of Clinical Epidemiology* 121 (May 2020), 20–28. doi:10.1016/j.jclinepi.2020.01.005

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science* 7, 2 (1983), 155–170.

Christian Gold, Moritz Körber, Christoph Hohenberger, David Lechner, and Klaus Bengler. 2015. Trust in Automation – Before and After the Experience of Take-over Scenarios in a Highly Automated Vehicle. *Procedia Manufacturing* 3 (2015), 3025–3032. doi:10.1016/j.promfg.2015.07.847

Pamela M. Greenwood, John K. Lenneman, and Carryl L. Baldwin. 2022. Advanced driver assistance systems (ADAS): Demographics, preferred sources of information, and accuracy of ADAS knowledge. *Transportation Research Part F: Traffic Psychology and Behaviour* 86 (April 2022), 131–150. doi:10.1016/j.trf.2021.08.006

Eddie Guo, Mehul Gupta, Jiawen Deng, Ye-Jean Park, Michael Paget, and Christopher Naugler. 2024. Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study. *Journal of Medical Internet Research* 26, 1 (Jan. 2024), e48996. doi:10.2196/48996 Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research.

Rayan Ebnali Harari, Kevin Hulme, Aliakbar Ebnali-Heidari, and Adel Mazloumi. 2019. How does training effect users' attitudes and skills needed for highly automated driving? *Transportation Research Part F: Traffic Psychology and Behaviour* 66 (Oct. 2019), 184–195. doi:10.1016/j.trf.2019.09.001

Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. doi:10.1016/S0166-4115(08)62386-9

Mary Hegarty, Mike Stieff, and Bonnie L. Dixon. 2013. Cognitive change in mental models with experience in the domain of organic chemistry. *Journal of Cognitive Psychology* 25, 2 (March 2013), 220–228. doi:10.1080/20445911.2012.725044

Frederic Hilkenmeier, Merle Stoltenberg, and Christian Stierle. 2026. Using full agreement across multiple large language models for title-and-abstract screening in systematic reviews: a proof-of-concept. *Systematic Reviews* 15, 1 (2026), 191. doi:10.1186/s13643-026-03228-4

Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57, 3 (May 2015), 407–434. doi:10.1177/0018720814547570

Yukio Horiguchi, Ryuichi Fukuju, and Tetsuo Sawaragi. 2006. An Estimation Method of Possible Mode Confusion in Human Work with Automated Control Systems. In *2006 SICE-ICASE International Joint Conference*. IEEE, Busan Exhibition & Convention Center-BEXCO, Busan, Korea, 943–948. doi:10.1109/SICE.2006.315649

Christian P. Janssen, Linda Ng Boyle, Andrew L. Kun, Wendy Ju, and Lewis L. Chuang. 2019. A Hidden Markov Framework to Capture Human–Machine Interaction in Automated Vehicles. *International Journal of Human–Computer Interaction* 35, 11 (July 2019), 947–955. doi:10.1080/10447318.2018.1561789 _eprint: https://doi.org/10.1080/10447318.2018.1561789.

James Jenness, John Lenneman, Amy Benedick, Richard Huey, Joshua Jaffe, Jeremiah Singer, and Sarah Yahoodik. 2019. Spaceship, guardian, coach: Drivers' mental models of advanced vehicle technology. In *International conference on human-computer interaction*. Springer, 351–356. doi:10.1007/978-3-030-23525-3_46

Philip Nicholas Johnson-Laird. 1983. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press. Number: 6.

Natalie A. Jones, Helen Ross, Timothy Lynam, Pascal Perez, and Anne Leitch. 2011. Mental Models: An Interdisciplinary Synthesis of Theory and Methods. *Ecology and Society* 16, 1 (2011). doi:10.5751/ES-03802-160146

Soyeon Kim, Fjollë Novakazi, and I.C. MariAnne Karlsson. 2025. Is conditionally automated driving a bad idea? Observations from an on-road study in automated vehicles with multiple levels of driving automation. *Applied Ergonomics* 129 (Nov. 2025), 104617. doi:10.1016/j.apergo.2025.104617

Soyeon Kim, Fjollë Novakazi, Elmer van Grondelle, René van Egmond, and Riender Happee. 2024. Who is performing the driving tasks after interventions? Investigating drivers' understanding of mode transition logic in automated vehicles. *Applied Ergonomics* 121 (Nov. 2024), 104369. doi:10.1016/j.apergo.2024.104369

Walter Kintsch. 1998. Comprehension : a paradigm for cognition. (1998).

Christina Kurpiers, Bianca Biebl, Julia Mejia Hernandez, and Florian Raisch. 2020. Mode Awareness and Automated Driving—What Is It and How Can It Be Measured? *Information* 11, 5 (May 2020), 277. doi:10.3390/info11050277

Minae Kwon, Malte F. Jung, and Ross A. Knepper. 2016. Human expectations of social robots. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Christchurch, New Zealand, 463–464. doi:10.1109/HRI.2016.7451807

Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and usability engineering group*. Springer, 63–76.

John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (March 2004), 50–80. doi:10.1518/hfes.46.1.50_30392

Patrick Lindemann, Tae-Young Lee, and Gerhard Rigoll. 2018. Supporting Driver Situation Awareness for Autonomous Urban Driving with an Augmented-Reality Windshield Display. In *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. 358–363. doi:10.1109/ISMAR-Adjunct.2018.00104

Yuan-Cheng Liu, Nikol Figalová, and Klaus Bengler. 2022. Transparency Assessment on Level 2 Automated Vehicle HMIs. *Information* 13, 10 (Oct. 2022), 489. doi:10.3390/info13100489

Ryuji Matsuo, Hailong Liu, Toshihiro Hiraoka, and Takahiro Wada. 2024. Enhancing the Driver's Comprehension of ADS's System Limitations: An HMI Providing Request-to-Intervene Trigger and Reason Explanation. In *2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS)*. IEEE, Toronto, ON, Canada, 1–7. doi:10.1109/ICHMS59971.2024.10555611

Mercedes-Benz Group. 2022. The front runner in automated driving and safety technologies. https://group.mercedes-benz.com/technology/autonomous-driving/driving/drive-pilot.html tex.howpublished: Corporate technology article.

Barbara Metz, Johanna Wörle, and Alexandra Neukum. 2024. An online-survey on user expectations and mental model of automated driving: The effects of automation description and technology readiness. doi:10.54941/ahfe1005214

Gary T Moore and Reginald G Golledge. 1976. *Environmental knowing: Theories, research and methods*. Dowden.

Michael A. Nees, Nithya Sharma, and Karli Herwig. 2020. Some Characteristics of Mental Models of Advanced Driver Assistance Systems: A Semi-structured Interviews Approach. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 64, 1 (Dec. 2020), 1313–1317. doi:10.1177/1071181320641314

Thomas O. Nelson. 1990. Metamemory: A Theoretical Framework and New Findings. In *Psychology of Learning and Motivation*. Vol. 26. Elsevier, 125–173. doi:10.1016/S0079-7421(08)60053-5

N. Neuhuber, P. Pretto, and B. Kubicek. 2022. Interaction strategies with advanced driver assistance systems. *Transportation Research Part F: Traffic Psychology and Behaviour* 88 (July 2022), 223–235. doi:10.1016/j.trf.2022.05.013

Khazar Dargahi Nobari and Torsten Bertram. 2025. Analysis of Driver Workload on Risk Perception in Non-Safety-Critical Situations with Partial Automation. In *2025 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. IEEE, Duisburg, Germany, 56–63. doi:10.1109/CogSIMA64436.2025.11079506

Sina Nordhoff, John D. Lee, Simeon C. Calvert, Siri Berge, Marjan Hagenzieker, and Riender Happee. 2023. (Mis-)use of standard Autopilot and Full Self-Driving (FSD) Beta: Results from interviews with users of Tesla's FSD Beta. *Frontiers in Psychology* 14 (Feb. 2023). doi:10.3389/fpsyg.2023.1101520

Donald A. Norman. 1983a. Design rules based on analyses of human error. *Commun. ACM* 26, 4 (April 1983), 254–258. doi:10.1145/2163.358092

Donald A. Norman. 1983b. Some Observations on Mental Models. In *Mental Models*. Psychology Press. Num Pages: 8.

Fjollë Novakazi. 2025. Perception's Role for Mental Model Formation in Automated Driving: Insights From Four Studies. *Journal of Cognitive Engineering and Decision Making* 19, 4 (Dec. 2025), 420–452. doi:10.1177/15553434251334409

Fjollë Novakazi, Alexander Eriksson, and Lars-Ola Bligård. 2022. Design for Perception - A Systematic Approach for the Design of Driving Automation Systems based on the Users' Perception. In *Adjunct Proceedings of the 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, Seoul Republic of Korea, 87–90. doi:10.1145/3544999.3552525

Fjollë Novakazi, Soyeon Kim, and I. C. MariAnne Karlsson. 2025. Mind Over Matter - Investigating the Influence of Driver's Perception in the Misuse of Automated Vehicles. In *Proceedings of the 17th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '25)*. Association for Computing Machinery, New York, NY, USA, 117–128. doi:10.1145/3744333.3747835

OpenAI. 2025. Introducing GPT-5.2. https://openai.com/index/introducing-gpt-5-2/ tex.howpublished: OpenAI product and evaluation report.

OpenAI. 2026. GPT-5.6: Frontier intelligence that scales with your ambition. https://openai.com/index/gpt-5-6/ tex.howpublished: OpenAI product and evaluation report.

Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Systematic Reviews* 10, 1 (March 2021), 89. doi:10.1186/s13643-021-01626-4

Raja Parasuraman and Victor Riley. 1997. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors* 39, 2 (June 1997), 230–253. doi:10.1518/001872097778543886

R. Parasuraman, T.B. Sheridan, and C.D. Wickens. 2000. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 30, 3 (May 2000), 286–297. doi:10.1109/3468.844354

Mickaël Jean Rémy Perrier, Tyron Linton Louw, and Oliver Carsten. 2021. User-centred design evaluation of symbols for adaptive cruise control (ACC) and lane-keeping assistance (LKA). *Cognition, Technology & Work* 23, 4 (Nov. 2021), 685–703. doi:10.1007/s10111-021-00673-0

Kristina Stojmenova Pečečnik. 2023. Situational awareness in the automotive domain - Effects on driving performance and driver's behavior. In *2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*. IEEE, Tenerife, Canary Islands, Spain, 1–5. doi:10.1109/ICECCME57830.2023.10253014

Linda Pipkorn, Trent W. Victor, Marco Dozza, and Emma Tivesten. 2021. Driver conflict response during supervised automation: Do hands on wheel matter? *Transportation Research Part F: Traffic Psychology and Behaviour* 76 (Jan. 2021), 14–25. doi:10.1016/j.trf.2020.10.001

Linwei Qiao, Jiaqian Li, and Tingru Zhang. 2024. Impacts of Training Methods and Experience Types on Drivers' Mental Models and Driving Performance. In *HCI in Mobility, Transport, and Automotive Systems*, Heidi Krömker (Ed.). Springer Nature Switzerland, Cham, 44–56. doi:10.1007/978-3-031-60477-5_4

Manuela Quaresma, Isabela Motta, Gabriel Martins, and Clara Gavinho. 2023. Evaluating Interface Layouts for Conditionally Automated Vehicle Messages. In *Design, User Experience, and Usability*, Aaron Marcus, Elizabeth Rosenzweig, and Marcelo M. Soares (Eds.). Springer Nature Switzerland, Cham, 284–303. doi:10.1007/978-3-031-35702-2_21

RDW. 2026. RDW explanation of european type approval for Tesla with provisional validity in the netherlands. https://www.rdw.nl/en/news/2026/rdw-explanation-of-european-type-approval-tesla-with-provisional-validity-in-the-netherlands tex.howpublished: News release.

Reuters. 2025. China's Xiaomi says it is cooperating with police after fatal EV accident. (April 2025). https://www.reuters.com/world/china/chinas-xiaomi-says-actively-cooperating-with-police-after-fatal-accident-2025-04-01/

Hansol Rheem, Joonbum Lee, John D. Lee, Joseph F. Szczerba, Roy Mathieu, and Akilesh Rajavenkatanarayanan. 2024. Investigating Drivers' Awareness of Automated Vehicle Modes Using the State Diagram Method. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 68, 1 (Sept. 2024), 894–898. doi:10.1177/10711813241275986

William B. Rouse and Nancy M. Morris. 1986. On looking into the black box: Prospects and limits in the search for mental models. *Psychological Bulletin* 100, 3 (1986), 349–363. doi:10.1037/0033-2909.100.3.349 Place: US.

J. Rushby, J. Crow, and E. Palmer. 1999. An automated method to detect potential mode confusions. In *Gateway to the New Millennium. 18th Digital Avionics Systems Conference. Proceedings (Cat. No.99CH37033)*, Vol. C.2-7 vol.1. IEEE, St Louis, MO, USA, 4.B.2–1–4.B.2–6. doi:10.1109/DASC.1999.863725

SAE International. 2021. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. https://www.sae.org/standards/j3016-taxonomy-definitions-terms-related-driving-automation-systems-road-motor-vehicles

David M. Sanbonmatsu, David L. Strayer, Zhenghui Yu, Francesco Biondi, and Joel M. Cooper. 2018. Cognitive underpinnings of beliefs and confidence in beliefs about fully automated vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour* 55 (May 2018), 114–122. doi:10.1016/j.trf.2018.02.029

Catarina Marques Santos, Ana Margarida Passos, and Sjir Uitdewilligen. 2016. When shared cognition leads to closed minds: Temporal mental models, team learning, adaptation and performance. *European Management Journal* 34, 3 (June 2016), 258–268. doi:10.1016/j.emj.2015.11.006

Nadine B. Sarter and David D. Woods. 1995. How in the World Did We Ever Get into That Mode? Mode Error and Awareness in Supervisory Control. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 37, 1 (March 1995), 5–19. doi:10.1518/001872095779049516

Dmitry Scherbakov, Nina Hubig, Vinita Jansari, Alexander Bakumenko, and Leslie A Lenert. 2025. The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. *Journal of the American Medical Informatics Association* 32, 6 (June 2025), 1071–1086. doi:10.1093/jamia/ocaf063

Sarah Schwindt-Drews, Kai Storms, Steven Peters, and Bettina Abendroth. 2024. Acceptance and Trust: Drivers' First Contact With Released Automated Vehicles in Naturalistic Traffic. *IEEE Transactions on Intelligent Transportation Systems* 25, 11 (Nov. 2024), 18601–18610. doi:10.1109/TITS.2024.3443927

Stephanie Seupke, Sarukan Segar, and Martin Baumann. 2025a. Development of a Mental Model Questionnaire Framework: A Systematic Approach to Measuring Mental Models in Automated Driving. doi:10.2139/ssrn.5168351

Stephanie Seupke, Sarukan Segar, and Martin Baumann. 2025b. Long-Term Evolution of Driver Visual Attention during Automated Driving in Real-Traffic: Investigating the Influence of Mental Model and Dynamic Learned Trust. In *Proceedings of the 17th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, Brisbane Australia, 318–329. doi:10.1145/3744333.3747821

Marina Strano, Fabricio Novak, Shelly Walbert, Beatriz Palmeiro, Sonia Morales, and Ignacio Alvarez. 2018. "Peace of Mind", An Experiential Safety Framework for Automated Driving Technology Interactions. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. 53–59. doi:10.1109/ITSC.2018.8569686

Wilbert Tabone, Joost C.F. de Winter, Claudia Ackermann, Jonas Bärgman, Martin Baumann, Shuchisnigdha Deb, Colleen Emmenegger, Azra Habibovic, Marjan Hagenzieker, P.A. Hancock, Riender Happee, Josef Krems, John D. Lee, Marieke H. Martens, Natasha Merat, Don Norman, Thomas B. Sheridan, and Neville A. Stanton. 2021. Vulnerable road users and the coming wave of automated vehicles: Expert perspectives. *Transportation Research Interdisciplinary Perspectives* 9 (March 2021), 100293. doi:10.1016/j.trip.2020.100293

Angelica M. Tinga, Diane Cleij, Reinier J. Jansen, Sander van der Kint, and Nicole van Nes. 2022. Human machine interface design for continuous support of mode awareness during automated driving: An online simulation. *Transportation Research Part F: Traffic Psychology and Behaviour* 87 (May 2022), 102–119. doi:10.1016/j.trf.2022.03.020

Angelica M. Tinga, Sebastiaan M. Petermeijer, Antoine J. C. de Reus, Reinier J. Jansen, and Boris M. van Waterschoot. 2023. Assessment of the cooperation between driver and vehicle automation: A framework. *Transportation Research Part F: Traffic Psychology and Behaviour* 95 (May 2023), 480–493. doi:10.1016/j.trf.2023.04.002

Trent W. Victor, Emma Tivesten, Pär Gustavsson, Joel Johansson, Fredrik Sangberg, and Mikael Ljung Aust. 2018. Automation Expectation Mismatch: Incorrect Prediction Despite Eyes on Threat and Hands on Wheel. *Human Factors* 60, 8 (Dec. 2018), 1095–1116. doi:10.1177/0018720818788164

Francesco Walker, Yannick Forster, Sebastian Hergeth, Johannes Kraus, William Payre, Philipp Wintersberger, and Marieke H. Martens. 2023. Trust in automated vehicles: constructs, psychological processes, and assessment. *Frontiers in Psychology* 14 (Nov. 2023). doi:10.3389/fpsyg.2023.1279271

Kyle M. Wilson, Shiyan Yang, Trey Roady, Jonny Kuo, and Michael G Lenné. 2020. Driver trust & mode confusion in an on-road study of level-2 automated vehicle technology. *Safety Science* 130 (Oct. 2020), 104845. doi:10.1016/j.ssci.2020.104845

Johanna Wörle and Barbara Metz. 2023. Misuse or abuse of automation? Exploring drivers' intentions to nap during automated driving. *Transportation Research Part F: Traffic Psychology and Behaviour* 99 (Nov. 2023), 460–472. doi:10.1016/j.trf.2023.10.023

Ji Hyun Yang, Jimin Han, and Jung-Mi Park. 2017. Toward Defining Driving Automation from a Human-Centered Perspective. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, Denver Colorado USA, 2248–2254. doi:10.1145/3027063.3053101

Jing Zang and Myounghoon Jeon. 2025. When more is less: Finding the optimal balance of intelligent agents' transparency in level 3 automated vehicles. *International Journal of Human-Computer Studies* 193 (Jan. 2025), 103384. doi:10.1016/j.ijhcs.2024.103384

Zhenzhen Zhuang, Jiandong Chen, Hongfeng Xu, Yuwen Jiang, and Jialiang Lin. 2025. Large language models for automated scholarly paper review: A survey. *Information Fusion* 124 (Dec. 2025), 103332. doi:10.1016/j.inffus.2025.103332