Human Interaction and Emerging Technologies (IHIET 2026), Vol. 236, 2026, 1–10
https://doi.org/10.54941/ahfe1008076

Explainability in Automated Driving: From Spatial Attention to Human-Centred Reasoning

Andry Rakotonirainy[1], Ashkan Y. Zadeh[1], Zishuo Zhu[1], Djamel Benrachou[1], Mohammed Elhenawy[1], Sebastien Glaser[1], Xiaomeng Li[1], Ronald Schroeter[1], Melaine Gouillou[2], and Patricia Delhomme[3]

[1]AVR3, CARRS-Q - Queensland University of Technology, Brisbane, Australia
[2]CentraleSupélec, Université Paris-Saclay, France
[3]Université Gustave Eiffel, France

ABSTRACT

Automated Vehicles (AVs) are developing rapidly and promise to improve road safety. AV systems are equipped with advanced AI techniques to perceive, learn, decide, and act, yet the decision-making process underpinned by complex AI constitutes a black box for human users. While human understanding of AI-based decision-making is critical to building trust, acceptance, and efficient human-machine cooperation, existing algorithmic approaches provide cause-effect relations in decisions but fail to account for the psychosocial and cognitive dynamic nature of supervising an AV. There is a lack of comprehensive framework addressing this algorithmic transparency with the cognitive and affective dynamics of the human recipient, leaving the human side of the explanation transaction theoretically underdeveloped. We propose a human-centred explainability framework that repositions AV explanation as a cognitive human-machine interaction problem. The framework is operating across spatial, semantic, and cognitive registers. It articulates where a system attends, what it recognises, and why that recognition contributes to better explanation. These levels integrate semantic awareness to enrich causal reasoning, while spatial information contextualise explanations. The preliminary framework is grounded in naturalistic field observations in instrumented vehicles, collecting video data and field notes complemented by post-drive semi-structured interviews. Thematic analysis yielded a driving context-specific explanatory vocabulary, formalising conditions under which users seek, process, and integrate AV explanations. Drawing on cognitive science and human factors theory, we derive design principles for adaptive explanation delivery, leveraging multimodal large language models as the generative engine for contextualised natural language explanations responsive to user expertise and situational urgency. This work demonstrates that the next frontier in AV explainability should be more aligned with human cognition.

Keywords: Automated vehicles, Explainability, Artificial intelligence, Human-machine interaction, Cognitive triggers

Received June 12, 2026; Revised June 22, 2026; Accepted July 10, 2026; Available online August 25, 2026

INTRODUCTION

SAE Level 3–5 Automated Vehicles (AVs) are transforming the driving task from active control to supervisory monitoring, yet this transition introduces a fundamental tension: the more capable the AV system, the opaquer its decision-making becomes to the human user expected to oversee it. AVs increasingly rely on advanced Artificial Intelligence (AI) and machine-learning models, including Deep Neural Networks (DNNs) deployed across perception, prediction, planning, and control. These models, often trained on high-dimensional, multimodal datasets, demonstrate strong capabilities in scene understanding and motion planning. Their internal computations, however, remain highly non-linear, distributed, and difficult to interpret, leaving them largely opaque to human reasoning. This opacity lies at the core of the trust, acceptance, and safety challenges that determine whether the societal promise of automated driving is realised

The field of eXplainable AI (XAI) has emerged to address this challenge, yet its dominant paradigms remain narrowly framed. The majority of XAI research is algorithmic-centric (Zhu et al., 2025; Zadeh et al., 2026; Atakishiyev et al., 2024): it focuses on making model internals visible through scoring systems, decision trees, or attribution methods, while treating the human recipient of explanations as a passive and homogeneous consumer. This framing is fundamentally inadequate for the AV context. Supervising an AV is a psychosocially and cognitively dynamic task, shaped by the driver's prior experience, risk perception, evolving trust, and self-regulated learning over time (Michon, 1985; Fuller, 2005).

Furthermore, an important dimension of explainability has yet to be properly addressed: the driver's attentional state at the moment a decision is made. A substantial body of research on driver distraction has established that gaze allocation and visual attention are primary determinants of hazard detection and crash risk. For instance, driver inattention is estimated to contribute to up to 23% of crashes and near-crashes (Regan, Lee, & Young, 2009). Yet this knowledge has not been articulated with research explainability research. The driver distraction field treats inattention as a problem to be detected and warned against; the XAI field treats explanation as a quality attribute of the system's decision. Neither field has asked whether explanation delivery should be conditioned on what the driver has and has not perceived during the AV acts.

Explanation relevance is determined not only by the nature of the AV's decision, but by what the driver has and has not seen — and by who the driver is. Drivers differ substantially in their capacity to process and integrate explanations, shaped by experience, expertise, and self-regulatory style (Michon, 1985; Fuller, 2005). Beyond capability, personal preferences for the depth, modality, and timing of explanations introduce a further layer of individual variation that a one-size-fits-all XAI system cannot accommodate. A genuinely human-centric approach must therefore answer not only what the system decided and why, but also when an explanation is needed and for whom.

Explanation need itself arises from two distinct forms of discrepancy: a mental model mismatch (Johnson-Laird, 1983; Endsley, 1995), where the AV's behaviour contradicts the driver's expectations of the system, and a cognitive schema violation (Bartlett, 1932), where an element of the driving

scene falls outside the driver's contextual expectations of the world. Both generate demands for explanation, but of qualitatively different kinds. This distinction matters for XAI HMI design.

To address these limitations, we propose a human-centred explainability framework that conceptualises explanation generation as a function of the discrepancy between the AV's knowledge state and the driver's perceptual, semantic, and cognitive state. As shown in Figure 1, the framework is organised around three interacting registers. The spatial register captures what the driver has perceived, the semantic register captures how driving scene elements are interpreted, and the cognitive register captures mental models, situational understanding, and explanation demand. Together, these registers support an explanation gating mechanism that determines when an explanation is needed, what information should be communicated, and how it should be adapted to the driver's state.

The remainder of this paper is organised as follows. Section 2 analyses the cognitive triggers that generate explanation demand. Section 3 addresses the conditionality of explanation — when, for whom, and under what attentional conditions explanations should be delivered. Section 4 presents two complementary empirical studies that operationalise the human-centred framework from the demand and supply sides respectively. Finally, Section 5 discusses the implications and outlines directions for future research.

HUMAN-CENTRED EXPLAINABILITY FRAMEWORK

Figure 1 presents the proposed human-centred explainability framework. The framework conceptualises explanation generation as a function of the discrepancy between the AV's internal state and the driver's perceptual, semantic, and cognitive state. It is framed around three critical and interacting registers: (1) Spatial Register, (2) Semantic Register, and (3) Cognitive Register. These registers progressively transform sensory information into human-understandable explanations.

1. Spatial Register: The spatial register describes what the driver has perceived within the driving scene. It captures gaze allocation, visual attention, fixation behaviour, and attentional gaps between the driver's visual awareness and the information available to the AV perception system. Its primary function is to determine whether relevant environmental cues have been observed by the driver.
2. Semantic Register: The semantic register describes the interpretation and significance of perceived scene elements. It captures how objects, events, and environmental conditions are understood in relation to driving goals, risk, and context. At this level, explanation moves beyond object detection toward meaning, including hazard interpretation, contextual relevance, and out-of-distribution event recognition.
3. Cognitive Register: The cognitive register describes how drivers integrate information into mental models of the vehicle, the driving environment, and the ongoing situation. It encompasses expectation formation, situation awareness, trust calibration, schema activation, and explanation demand. Discrepancies within this register give rise to

mental model mismatches and schema violations that trigger requests for explanation.

4. Explanation Gating: The three registers jointly support an explanation gating mechanism that determines whether an explanation is necessary, what content should be communicated, and how that content should be adapted to the driver's attentional state, expertise, situational awareness, and cognitive workload.

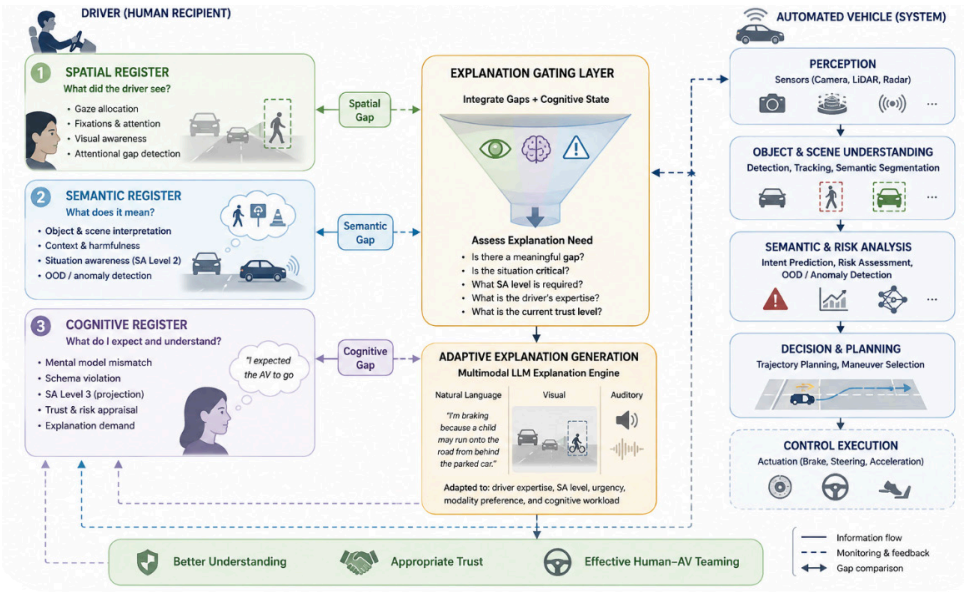

Figure 1: Human centred explainability for automated driving: Explanations are triggered by discrepancies between AV's knowledge state and the driver's perceptual, semantic and cognitive state.

COGNITIVE TRIGGERS OF EXPLANATION DEMAND

The cognitive register shown in Figure 1 is activated when discrepancies arise between the driver's expectations and either the AV's behaviour or the surrounding driving environment. Such discrepancies constitute the primary triggers of explanation demand. This section identifies three distinct trigger types: mental model mismatch, cognitive schema violation, and out-of-distribution hazard events. Each generates a qualitatively different explanatory need and has distinct implications for XAI HMI design.

Mental Model Mismatch

A mental model mismatch occurs when the AV's behaviour conflicts with the driver's expectations of the system (Johnson-Laird, 1983; Endsley, 1995). Humans seek explanations selectively because they serve as a mechanism for reconciling contradictions between expectations and observed behaviour. Self-regulated learning modulates this demand: a driver who has encountered

a behaviour repeatedly will progressively internalise it and cease to require explanation, while novel or high-risk situations will re-activate explanation-seeking regardless of prior exposure (Michon, 1985). An XAI system blind to these dynamics will over-explain in familiar situations, generating unnecessary cognitive load, and under-explain in novel ones, leaving the user without the understanding needed to maintain appropriate situational awareness.

Cognitive Schema Violation

A second, theoretically distinct trigger occurs when the driver has seen an object but has not correctly interpreted it, because it falls outside their cognitive schema for the driving context (Bartlett, 1932). A driver who fixates on a skateboard in a school zone, a mattress on a motorway, or a horse crossing an intersection has not failed at the perceptual level — they have encountered a schema-discrepant event whose meaning and implications for the AV's response they cannot readily assimilate. In cognitive terms, this constitutes a violation of expectation: an event that generates a qualitatively different kind of explanatory demand, one concerned not with awareness solely but with comprehension and projection, as reflected in the second and third levels of Endsley's Situation Awareness model (1995).

Unexpected, schema-discrepant events cause an automatic interruption of ongoing mental processes, followed by an attentional shift, causal analysis, and schema revision (Reisenzein et al., 2019; Retell et al., 2016). This has direct implications for XAI HMI design: a schema-violating event requires the AV to provide a causal account of why its response to that object is what it is, at a moment when the driver's cognitive system has already been disrupted. Delivering a poorly timed or cognitively overloading explanation here risks compounding the comprehension failure rather than resolving it.

Out-of-Distribution Hazard Detection

On the AI side, schema violation maps onto the well-recognised problem of OOD object detection. Conventional DNN-based methods classify objects into a fixed set of known classes, which limits the AV's ability to respond appropriately to objects outside that set. Recent work sidesteps this constraint by shifting emphasis from classification toward harmfulness determination, inferring risk from an object's position relative to the vehicle's trajectory instead of from its identity. This is enough for the vehicle to act: it can brake or swerve for an object it cannot name.

The same independence from object identity that makes this approach effective for action is what makes it inadequate for explanation. Detecting that an OOD object is present and harmful is necessary but not sufficient: the driver needs to understand not only that the AV has reacted, but what the anomalous object is and why it triggered that reaction, in language that maps onto their schema-revision process, not a classification taxonomy (Avetisyan et al., 2025).

A schema-discrepancy signal — computed by comparing the AV's object classification confidence against a contextual prior of expected road scene elements — could serve as an independent XAI trigger regardless of whether

the driver has fixated on the object. Research on designing explanations around Situation Awareness (SA) levels shows that SA Level 2 explanations, which address comprehension of what detected objects mean for AV behaviour, produce the highest situational trust, even at the cost of increased mental workload (Avetisyan et al., 2025). Integrating schema-discrepancy detection into the explanation gating architecture is a substantive challenge at the intersection of anomaly detection, cognitive modelling, and adaptive HMI design.

CONDITIONALITY OF EXPLANATION DELIVERY

The conditionality of XAI, such as when and for whom an explanation is necessary, is an emerging but under theorised area. Drivers do not need explanations continuously, and that presenting explanations only in critical driving conditions is preferred, precisely because indiscriminate explanation generation increases cognitive load and disrupts the driving experience (Kim et al., 2023). This is consistent with the broader XAI and AV literature's observation that the "four W" questions. Who needs an explanation, why, what kind, and when remain largely unresolved, with temporal conditionality receiving the least systematic attention (Atakishiyev et al., 2024).

The Attentional Gap as a Trigger

Existing frameworks address explanation timing by detecting unexpected AV behaviour or heightened crash risk. They do not account for whether the driver has independently perceived the relevant hazard cues. Consider the case of an AV decelerating in response to a pedestrian about to cross. If the driver has already detected and fixated on that pedestrian, an explanation is redundant and risks adding cognitive load. If the driver has not perceived the pedestrian, the same explanation becomes safety-critical. This bridges the gap between the AV's perception and the human's, enabling the driver to understand and, if necessary, endorse or override the system's action.

Closing such a gap requires coupling explanation delivery with real-time driver attention pattern. This approach is increasingly tractable with the use of advances in gaze-based situational awareness prediction (Lim et al., 2025) and by frameworks that align machine and human visual attention to establish common ground for human-machine interaction. The use of Theory of Mind in XAI has formalised the underlying logic: genuinely informative explanations bridge the gap between the machine's knowledge state and the human's inferred knowledge state, rather than restating what the user already knows (Vedantam et al., 2022). Applying this to the attentional domain by using predicted driver gaze as a proxy for what the driver has perceived would allow AV explanation systems to operate on a relevance criterion rather than a decision criterion.

Driver Capability, Self-Regulation, and Preference

Attentional state alone is not sufficient as a basis for adaptation. Drivers differ substantially in their capacity to process and integrate explanations, shaped

by experience, expertise, and self-regulatory style (Michon, 1985; Fuller, 2005). A novice driver encountering an unfamiliar scenario may require a fuller causal account of the AV's action, while an experienced driver may need only a brief alerting cue. Beyond capability, personal preferences for the depth, modality, and timing of explanations introduce a further layer of individual variation that a one-size-fits-all XAI system cannot accommodate. Effective adaptive explainability must therefore operate across three intersecting dimensions simultaneously. What the driver has perceived, what they are capable of processing given their expertise and current cognitive load, and what form of explanation they prefer and trust are complex to implement.

This shifts the central question of AV explainability from "how do we make AI decisions interpretable?" to "how do we make AI explanations relevant?". This is a distinction with profound implications for interface design, cognitive load management, and road safety. Realising it will require tighter integration between driver monitoring systems, attention prediction models, and natural language explanation engines. It raises important open questions about the appropriate latency, modality, and granularity of conditionally triggered explanations in safety-critical contexts.

EMPIRICAL STUDIES TOWARDS A HUMAN-CENTRED XAI FRAMEWORK

The theoretical framework advanced in previous sections raises two distinct empirical questions: what explanation structures do human drivers actually produce in naturalistic driving contexts (the demand side), and how should those structures be linguistically formulated for cognitive accessibility (the supply side)? The two studies presented below attempt to answer each question. Together bracketing the human-centred XAI problem from both ends.

Study 1: Verbal Protocol Analysis of Driver Explanation Behaviour

The cognitive register concerned with how human recipients in AV seek, interpret, and integrate AV explanations remains the least empirically grounded. A published systematic review established the empirical landscape across 59 peer-reviewed studies (Zhu et al., 2025). Existing work has characterised what information users prefer and when they prefer to receive it, yet little is known about the reasoning structures that underpin those preferences, or how explanation content should be derived from documented human expertise rather than developer assumption (Omeiza et al., 2021). A key finding was that existing AV explanation approaches predominantly rely on developer-defined content with little user involvement, and that explanations are rarely tailored to specific driving contexts — underscoring the need for empirically grounded, context-sensitive explanation design.

Building on this, an empirical study applies a cognitive framework to analyse verbal protocols from a naturalistic driving dataset of N = 37 experienced drivers, including driving instructors. Verbal protocol analysis

has an established tradition in capturing expert reasoning in complex, real-world tasks (Ericsson, 1998), and has been applied in driving contexts to surface the cognitive processes underlying skilled performance. Initial findings indicate that expert drivers employ different explanation patterns across driving contexts, which are being formalised as candidate explanation templates. These represent a human-derived content specification that grounds the cognitive register of the framework in documented expert reasoning, rather than in developer assumption or model output alone — addressing calls for explanation content to reflect authentic user-centric design (Gyevnar et al., 2025; Zhu et al., 2025). A subsequent user study will evaluate these templates, examining how AV passengers seek and process explanations over the course of a continuous automated drive.

Study 2: Linguistic and Scenario-Sensitive Explanation Generation

Psycholinguistics provides a critical mechanism for explaining how people process language under cognitive constraints. In automated driving, explanations are often received under divided attention, uncertainty, and time pressure. Conditions under which technically correct explanations may still fail if they are syntactically complex or poorly aligned with user expectations. Our PsyLingXAV framework addresses this by integrating psycholinguistic principles into AV explanation design, proposing that raw decision outputs be refined through a psycholinguistic layer before delivery (Zadeh et al., 2025). This layer supports rapid comprehension through simpler sentence structures, avoidance of garden-path ambiguity, alignment with user expectations through cloze probability, and syntactic priming through familiar sentence patterns.

However, human-centred reasoning also requires explanations to be sensitive to driving scenario. The framework addresses this by providing a scenario-aware linguistic knowledge acquisition system. Rather than relying on generic language generation. It extracts structured knowledge from human-authored driving explanations across three levels: scenario context, syntactic structure, and lexical choice, producing a hierarchical linguistic knowledge base that guides the design and evaluation of AV explanation systems.

The framework provide the linguistic bridge between machine-centred perception and human-centred reasoning. It determines what should be explained in a given context and which linguistic resources are appropriate; PsyLingXAV determines how the explanation should be phrased for efficient user processing. The framework integrates semantic interpretation, cognitive explanation and constrains multimodal Large Language Model (LLM)-based generation to produce outputs that are contextually grounded, cognitively manageable, and aligned with authentic driving explanation patterns.

DISCUSSION AND FUTURE RESEARCH

The two empirical studies presented in this paper operationalise complementary aspects of the human-centred XAI framework. Study 1 establishes, from verbal

protocol evidence, the type of explanation structures expert drivers naturally produce — providing a demand-side specification grounded in human cognition rather than developer assumption. Study 2 provides the supply-side architecture: the psycholinguistic constraints that ensure generated explanations are both scenario-appropriate and cognitively accessible. Together, they demonstrate that designing for the human recipient of explanations — rather than for the algorithmic source — is both necessary and tractable.

Three directions represent the most substantive open challenges for future research. The first is the integration of real-time driver attention modelling into explanation gating: pairing DMS gaze data with the AV's semantic object detection to compute a perception gap that drives explanation triggers, instead of relying on system decision events alone. This requires tighter integration between driver monitoring systems, attention prediction models, and natural language explanation engines such as those built on LLMs, and raises open questions about the appropriate latency, modality, and granularity of conditionally triggered explanations. The second is the extension of the schema-discrepancy signal to operationalise OOD-triggered explanations: building a contextual prior of expected scene elements against which anomalous objects can be detected and communicated. The third is longitudinal validation of trust dynamics under sustained explanation exposure, together with extension of the framework to edge cases such as system failures and ethical dilemmas, where explanation demands are highest and current approaches most severely inadequate.

The unifying design principle should be that explanation relevance is determined not only by what the system has done, but by the gap between what the system knows and what the individual driver has perceived, expect, and understood, and is capable of integrating in real time.

CONCLUSION

This paper has argued that the next frontier in AV explainability lies not in more interpretable models, but in more cognitively aligned explanation systems. By repositioning explainability as a cognitive human-machine interaction problem, and grounding the preliminary results in naturalistic empirical evidence on both the demand and supply sides of explanation, this work establishes a theoretically grounded and empirically principled basis for adaptive XAI in automated driving. The framework operates across spatial, semantic, and cognitive registers, conditioned on driver attention and capability, and sensitive to the cognitive triggers that generate explanation demand. It provides a foundation for explanation systems that communicate not just what the AV has done, but what the driver needs to know, when they need to know it, and in a form they can act upon.

ACKNOWLEDGMENT

The authors would like to acknowledge Australian Research Council grant support (DP220102598).

REFERENCES

Akula, A. R., Wang, K., Liu, C., Saba-Sadiya, S., Lu, H., Todorovic, S., Chai, J., & Zhu, S.-C. (2022). CX-ToM: Counterfactual explanations with theory-of-mind for enhancing human trust in image recognition models. iScience, 25(1).

Atakishiyev, S., Salameh, M., Yao, H., & Goebel, R. (2024). Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions. IEEE Access.

Avetisyan, L., Yang, X. J., & Zhou, F. (2025). Towards Context-Aware Modeling of Situation Awareness in Conditionally Automated Driving. International Journal of Human–Computer Interaction, 42(1), 291–306.

Bartlett, F. C. (1932). Remembering: A study in experimental and social psychology. Cambridge University Press.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. Human Factors, 37(1), 32–64.

Ericsson, K. A. (1998). Protocol Analysis. In W. Bechtel & G. Graham (Eds.), A Companion to Cognitive Science (pp. 425–432). Wiley Blackwell.

Fuller, R. (2005). Towards a general theory of driver behaviour. Accident Analysis & Prevention, 37(3), 461–472.

Gyevnar, B., Droop, S., Quillien, T., Cohen, S. B., Bramley, N. R., Lucas, C. G., & Albrecht, S. V. (2025). People Attribute Purpose to Autonomous Vehicles When Explaining Their Behavior. Conference on Human Factors in Computing Systems. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems.

Johnson-Laird, P. N. (1983). Mental models: Towards a cognitive science of language, inference, and consciousness. Cambridge University Press.

Kim, G., Yeo, D., Jo, T., Rus, D., & Kim, S. (2023). What and when to explain? On-road evaluation of explanations in highly automated vehicles. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 7(3), 1–26.

Lim, C., Liang, N., Villarreal, R. T., Prakah-Asante, K. O., Pitts, B. J., & Yu, D. (2025). Multimodal prediction of situation awareness during automated driving: a gaze and EEG-based approach. Ergonomics, 1–30.

Michon, J. A. (1985). A critical view of driver behavior models. In L. Evans & R. C. Schwing (Eds.), Human behavior and traffic safety (pp. 485–520). Plenum Press.

Omeiza, D., Kollnig, K., Web, H., Jirotka, M., & Kunze, L. (2021). Why Not Explain? Effects of Explanations on Human Perceptions of Autonomous Driving. Proceedings of IEEE Workshop on Advanced Robotics and Its Social Impacts, 194–199.

Regan, M. A., Lee, J. D., & Young, K. (Eds.). (2009). Driver distraction: Theory, effects, and mitigation. CRC Press.

Reisenzein, R., Horstmann, G., & Schützwöhl, A. (2019). The Cognitive-Evolutionary Model of Surprise: A Review of the Evidence. Topics in Cognitive Science, 11(1), 50–74.

Retell, J.D., Becker, S.I. & Remington, R.W (2016). Previously seen and expected stimuli elicit surprise in the context of visual search. Atten Percept Psychophys 78, 774–788.

Zadeh AY, Li X, Rakotonirainy A, Schroeter R, Glaser S, Zhu Z (2026). X-Blocks: Linguistic Building Blocks of Natural Language Explanations for Automated Vehicles. arXiv preprint arXiv:2602.13248. 2026.

Zhu, Z., Li, X., Delhomme, P., Schroeter, R., Glaser, S., & Rakotonirainy, A. (2025). Human-Centric explanations for users in automated Vehicles: A systematic review. Accident Analysis & Prevention, 220, 108152.