Experts' and Community Insights on Measuring Comfort and Perceived Safety in Automated Driving: Synthesis of a Participatory Workshop

CHEN PENG✉, Transport Safety Research Centre, School of Design and Creative Arts, Loughborough University , UK
PAVLO BAZILINSKYY, Department of Industrial Design, Eindhoven University of Technology, Netherlands
YUETENG YU, Queensland University of Technology, Australia
MARIEKE MARTENS, Department of Industrial Design, Eindhoven University of Technology, Netherlands
RIENDER HAPPEE, Department of Cognitive Robotics, Faculty of Mechanical Engineering, Delft University of Technology, Netherlands
KUMAR AKASH, Human and Social Sciences, Honda Research Institute USA, Inc., USA
NATASHA MERAT, Institute for Transport Studies, University of Leeds, UK

Accurately measuring comfort and perceived safety in automated vehicles (AVs) is critical for understanding user experience, accommodating various needs, and eventually contributing to the public acceptance of AVs. However, methodological practices vary largely across the field. To address this gap, we organised a workshop at AutomotiveUI 2025, where three expert talks covered subjective measurements, objective and physiological measures, and AI-driven models, followed by two structured group discussion sessions on current practices and future methodologies. This work-in-progress synthesises key insights from both the expert talks and community discussions. Critical themes include, for example, the need for context-specific, deliberate selection of subjective measures; the promise and practical limitations of physiological signals; the ground truth challenge that complicates subjective-objective combination; and the emerging role of artificial intelligence (AI) for real-time inference and multimodal combination. We offer these insights to provide the community with a common methodological landscape and to promote more robust practices.

CCS Concepts: • **Human-centered computing → User studies**.

ACM Reference Format:
Chen Peng, Pavlo Bazilinskyy, Yueteng Yu, Marieke Martens, Riender Happee, Kumar Akash, and Natasha Merat. 2026. Experts' and Community Insights on Measuring Comfort and Perceived Safety in Automated Driving: Synthesis of a Participatory Workshop. In *18th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI Adjunct '26), September 20–23, 2026, Gothenburg, Sweden.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3828158.3838256

Authors' Contact Information: Chen Peng, c.peng@lboro.ac.uk, Transport Safety Research Centre, School of Design and Creative Arts, Loughborough University, Loughborough, UK; Pavlo Bazilinskyy, Department of Industrial Design, Eindhoven University of Technology, Eindhoven, Netherlands, p.bazilinskyy@tue.nl; Yueteng Yu, Queensland University of Technology, Brisbane, Australia, yueteng.yu@hdr.qut.edu.au; Marieke Martens, Department of Industrial Design, Eindhoven University of Technology, Eindhoven, Netherlands, m.h.martens@tue.nl; Riender Happee, Department of Cognitive Robotics, Faculty of Mechanical Engineering, Delft University of Technology, Delft, Netherlands, R.Happee@tudelft.nl; Kumar Akash, Human and Social Sciences, Honda Research Institute USA, Inc., San Jose, CA, USA, kakash@honda-ri.com; Natasha Merat, Institute for Transport Studies, University of Leeds, Leeds, UK, n.merat@its.leeds.ac.uk.

1 Background

Comfort and perceived safety are two closely related yet conceptually different psychological states shaping user experience in automated vehicles (AVs), with both affecting public acceptance and adoption of AVs [3, 7, 10, 13]. Despite their importance, measuring these states presents substantial challenges. They are subjective, highly individual, and context-dependent [12]. They overlap with other constructs such as trust and acceptance [5, 11, 13]. Moreover, existing measurement approaches each have certain limitations. Subjective measures (e.g., Likert scales) lack cross-study comparability, and objective physiological signals are often susceptible to artefacts and individual variation [1, 12]. The emerging AI-based approaches remain largely underexplored for measurement purposes. Reliable measurement is critical for understanding users' comfort and perceived safety, developing robust indicators, and informing adaptive AV designs. To address the lack of shared knowledge on how to effectively measure these psychological states, we organised a workshop at the 17th International ACM Conference on Automotive Interfaces and Interactive Vehicular Applications (AutomotiveUI 2025) in Brisbane, Australia [12]. Three invited talks on different perspectives were provided, followed by two structured group discussion sessions. This Work-in-Progress (WiP) reports the key insights and community perspectives emerging from the workshop, with the aim of providing the AutomotiveUI community with a shared methodological landscape to guide future research and practice.

2 Method

2.1 Workshop structure

The workshop started with an introduction session to the workshop by the first author, followed by three expert talks, and finally two structured group discussion sessions. The expert talks covered 1) conceptual grounding of perceived safety and subjective measurement methods; 2) objective, physiological, and model-based measures; and 3) AI-driven real-time sensing pipelines for passenger state detection. The first discussion session focused on participants' past experiences with subjective and objective measures, such as instrument selection, lessons learned, and challenges encountered. The second discussion session focused on forward-looking strategies for multimodal measurement, including the potential role of artificial intelligence (AI) and/or large language models (LLMs) for real-time inference. Of the more than 40 attendees, approximately 22 participated in the discussion sessions, which were divided into four groups with the organisers as facilitators. Attendants wrote down their insights briefly on post-it notes and discussed each topic in sequence. All participants were informed that their contributions would be used for publication to share with the wider community and were free to withdraw at any time.

2.2 Approach to synthesising insights

Regarding the experts' talks, the main moderator wrote a summary of each presentation, and the experts had a chance to provide edits and feedback for the summary to ensure it correctly reflected the essence of their talks. Regarding the discussion sessions the written notes were transcribed and categorised based on their thematic features. This follows the procedure of reflective thematic analysis [2, 15], where written notes with similar semantic meanings were categorised together. Due to the ambiguity of some notes, 14 out of 83 notes from session 1, and 11 out of 56 notes from session 2 were not categorised. A full list, including original written notes, is available at (link removed for review). A generative AI tool (Claude Sonnet 4.6) was used to assist with digitising handwritten notes, and all outputs were manually checked and corrected by the first author. It was also used as a reflective tool to support the categorisation process.

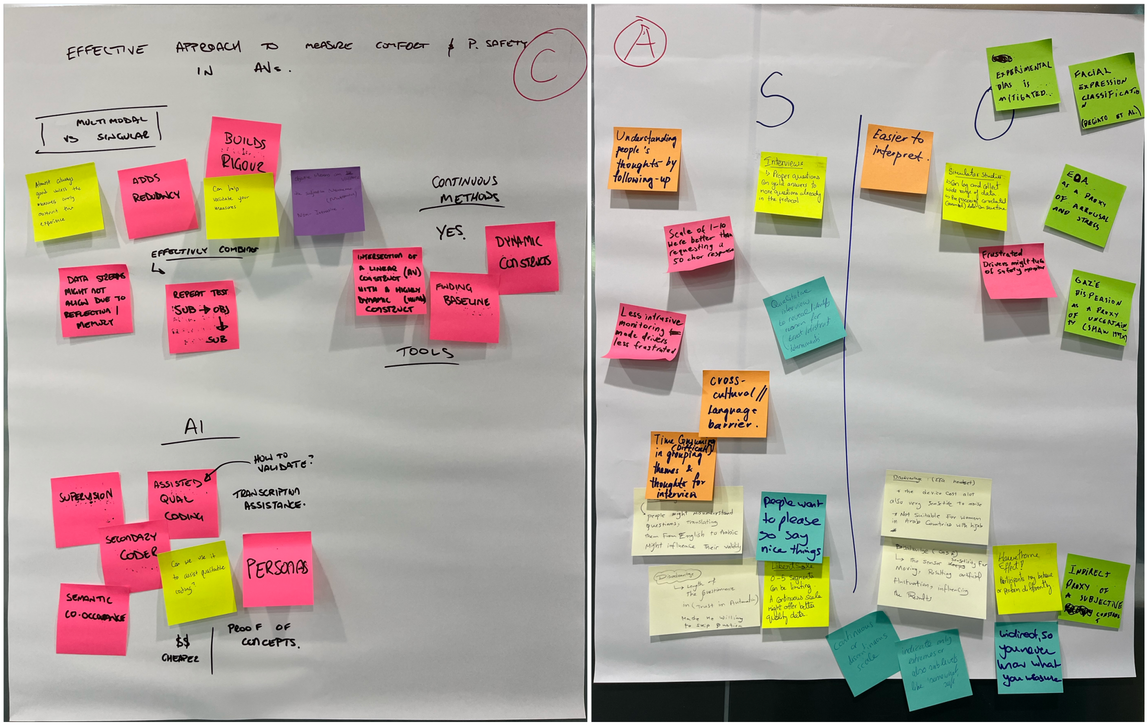

Fig. 1. Example of the output from the discussion sessions. The rest can be found in the supplementary file (link removed for review).

3 Experts' perspectives

3.1 Topic 1: Subjective measures

Perceived safety was characterised as subjective, highly individual, non-binary, and dynamic. It changes over time, context, and experience, and is linked to emotions such as fear, relaxation, and excitement rather than being a simple binary state. The concept is embedded in a dense network of related constructs, including trust, uncertainty, expectations, acceptance, perceived control, comfort, and reliability, suggesting that the high intercorrelation adds risks in measurement. Scales are often defined with simple instructions like rate "perceived safety" on a scale from 1-10. This leaves ample room for interpretation, which may hamper interpretation and induce major variance between respondents. This can be resolved by labelling several levels like "Very unsafe / Safe / Very safe". Such labelling could describe the perceived safety of the AV user, but could also capture social acceptability of AV behaviour for other users with labels such as "Aggressive / Socially acceptable / Overly compliant". Labels can also link to user behaviour or acceptance, like "I would overrule automation / I would be alert and actively monitor / I would take my eyes off the road". Key recommendations for subjective measurement included: anchoring items to a specific context and scenario rather than asking general questions; collecting relevant participant background (e.g., technology optimism, AV experience) as potential moderators; and including a baseline condition to assist interpretation. Continuous measures in time, using sliders, rotary knobs, button-press, offer temporal sensitivity but introduce practical challenges, including response delay, scale directionality (e.g., whether up or down corresponds to safer), compliance failures in naturalistic settings, and the need for manual reset between trials. Issues of button pressing include the difficulty in distinguishing variations within participants and the possibility of too few pressing points (due to many reasons, e.g., feeling unsafe, distracted, or misunderstanding the instruction). A continuous rotary knob paradigm produced clear event-linked response profiles in a simulator but suffered compliance failures on the road, where users prioritised monitoring the vehicle over reporting perceived safety [5, 14]. The overall recommendation was to treat all measures as imperfect and select and combine them deliberately.

3.2 Topic 2: Objective measures and computational models

Drawing on simulator and on-road studies, key findings include: skin conductance, also known as electrodermal activity (EDA), was one of the most sensitive physiological correlate of perceived safety, responding reliably to driving events; pupil diameter showed moderate correlations at critical moments [5]; heart rate increased during events but did not correlate with criticality; and facial emotion recognition largely failed, as the majority of participants showed neutral facial expressions at safety-critical events, with fear being the least detected emotion [4]. These findings highlight the need for calibration using known low- and high-risk events, given substantial individual variation. Computational models predicting perceived risk from vehicle kinematics are shown to well match subjective results in controlled simulator experiments. Human-machine Interfaces (HMIs) communicating the AV's perception and intent were shown to enhance perceived safety [6], which suggests that measurement, modelling, and interface design are deeply interrelated.

3.3 Topic 3: AI-driven models

This topic focused on a progression, moving from self-reports (closest to ground truth but intrusive and infeasible for real-time use) through physiological correlates to continuous AI models capable of sensing passenger states continuously and informing system behaviour. Perceived safety is positioned as the foundational layer of a hierarchy of passenger states, with trust and prosocial wellbeing building above it. A central insight was that labelling and windowing are the hardest part of building AI sensing models: the Selective Windowing Attention Network (SWAN) addressed this by learning which temporal windows matter, automatically identifying safety-critical moments rather than assuming fixed windows [9]. To address the small-dataset problem, self-supervised pretraining across 288 participants from multiple public datasets, followed by fine-tuning on a smaller target dataset, achieved 79.2% accuracy in predicting prosocial behaviour [8]. Key practical recommendations included: aligning labels with events, using subject-independent data splits for generalisation, reporting F1 and AUC (Area Under the Curve) under class imbalance, and treating measurement as a closed loop that feeds back into system behaviour.

4 Community perspectives

4.1 Good practices and challenges

4.1.1 Subjective measures. Participants discussed a range of approaches in designing and selecting subjective measures. Instrument selection should be grounded in the literature, with preference given to validated, widely cited scales. Regarding scale format, numerical rating scales were seen as more practical than open-ended responses, though standard Likert scales (e.g., 0-5 segments) can be limiting. Whether continuous or discrete scales are preferable remains unclear. Retrospective reporting afterwards was seen as more feasible, but it may be affected by memory. Qualitative and think-aloud methods were valued as complements to quantitative scales, especially for exploring reasons behind trust or distrust, although additional effort in analysis is introduced.

Table 1. Themes and categories of notes derived from the discussion session 1

Theme	*Category of notes*
Choices & good practices (Sub)	Instrument selection
	Scale format and granularity
	Timing of administration
	Think-aloud and qualitative options
Validity threats & challenges (Sub)	Social desirability & different behaviour
	Language and cultural barriers
	Questionnaire fatigue and usability
	Confounds
Selection and practical considerations (Ob)	Physiological signals
	Behavioural signals
	Practical constraints
	Study context/setting
Linking subjective and objective measures	Multimodal approaches
	Ground truth challenge

Threats to the validity were identified. Social desirability emerged as a concern, as participants may give overly positive responses to please the researchers or behave differently when being observed (Hawthorne Effect). Language and cultural barriers pose a further challenge, particularly when translating established scales across languages, where interpretation or response may change. Questionnaire fatigue and poor usability were also noted, with long surveys or items that are difficult to understand may lead to unreliable responses.

4.1.2 Objective measures. Physiological signals (e.g., EDA, EEG, ECG, EMG, facial expression, and thermal imaging) and observable behaviours (e.g., brake hovering, steering wheel contact) were identified as valuable proxies for subjective states. However, practical constraints were raised. Many physiological devices are expensive, sensitive to movement artefacts, or culturally unsuitable (e.g., EEG with hijab-wearing participants), and using too many objective measures simultaneously can feel obtrusive. Simulator settings can provide controlled data collection, while real-world and field settings introduce noise and more complexities, but add ecological validity.

4.1.3 Linking subjective and objective measures. Combining subjective and objective measures is valued but raises a fundamental ground truth problem. Since objective signals are indirect proxies, it is often unclear what is actually being measured or whether a stable baseline can be established. This uncertainty extends to AI-based inference. The question of how emotions or cognitive states manifest in measurable signals remains a conceptual challenge.

4.2 Discovering future methodologies

4.2.1 Measurement Approach. Participants broadly favoured multimodal measurement, suggesting that combining methods adds redundancy, builds rigour, and allows objective measures to validate subjective ones. Subjective ratings can be used to "scale" objective measures and compensate for individual variation, in particular for physiological measures. However, instrument combination raises practical challenges around synchronisation and data stream misalignment, particularly when subjective measures rely on retrospective recall. Think-aloud protocols and naturalistic data were noted as potentially problematic to analyse due to noise and the disruptive nature of verbalisation. A sequenced approach (subjective → objective → subjective) was proposed as one way to enhance measure combination. On

Table 2. Themes and categories of notes derived from the discussion session 2

Theme	*Category of notes*
Measurement approach	Multimodal measurement
	Instrument combination and challenges
	Sampling frequency
	Physiological/behavioural tools
	Ground truth and baseline
Individual differences	Individual difference
	Optimisation for same user
Role of AI	AI for real-time inference
	AI as qualitative analysis tool
	Conversational AI as a probe

sampling, (near)continuous measurement in time was seen as preferable given the dynamic nature of comfort and perceived safety across an AV journey. However, the appropriate sampling frequency remains an open question. Probing subjective rating after individual events creates specific insight and enhances the statistical significance of effects. Establishing ground truth and stable baselines was highlighted as an overarching challenge.

4.2.2 Individual Differences. Participants agreed that comfort and perceived safety are highly individual, varying across demographics, cultures, and personal baselines. It poses challenges for questionnaire design and generalisation. Balancing the preferences of multiple users was also raised as a particular challenge. Two approaches for single-user optimisation were discussed: allowing users to configure their own measurement settings and profiles, and building systems that learn incrementally from the same individual through feedback loops.

4.2.3 Role of AI. Participants identified multiple roles for AI in this context. For real-time inference, fine-tuned vision-language models were noted as capable of detecting facial features and signs of distracted driving, and AI could continuously record in-cabin conversations. For qualitative analysis, AI was discussed as a potential secondary coder for transcription and semantic analysis, although the validation of AI-assisted coding remained unclear. Conversational AI, as an active probe, engages participants in dialogue during a drive as a secondary task, which could capture users' experiences with low effort.

5 Discussion and Conclusion

The expert talks and community discussions revealed several shared themes. No single measurement method is perfect and sufficient; a deliberate multimodal combination is valuable; and the ground truth problem (i.e., what comfort and perceived safety actually are and how they manifest in measurable signals) remains a central challenge. Experts shared concrete guidance on signal selection and model design, while community discussions explored practical barriers, such as cultural constraints and questionnaire fatigue. The role of AI was explored, including real-time inference and an analytical assistant for qualitative data.

This work has limitations. The thematic analysis of post-it notes was exploratory rather than systematic, and some notes were difficult to interpret without additional context from the original discussion. Future work should focus on developing and validating multimodal measurement protocols with particular attention to ecological validity, individual differences, and ground-truth establishment.

Acknowledgments

We would like to thank the workshop participants' for their engagement and discussions.

References

[1] Matthias Beggiato, Franziska Hartwich, and Josef Krems. 2019. Physiological correlates of discomfort in automated driving. *Transportation Research Part F: Traffic Psychology and Behaviour* 66 (2019), 445–458. TLDR: The research project KomfoPilot at Chemnitz University of Technology aimed to investigate factors that influence comfort in automated driving as well as to identify physiological indicators of discomfort to provide a basis for developing a real-time discomfort detection algorithm.. doi:10.1016/j.trf.2019.09.018

[2] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. doi:10.1191/1478088706qp063oa

[3] Oliver Carsten and Marieke H. Martens. 2018. How can humans understand their automated cars? HMI principles, problems and solutions. *Cognition, Technology and Work* 21, 1 (2018), 3–20. ISBN: 0123456789. doi:10.1007/s10111-018-0484-0

[4] Abel van Elburg, Konstantinos Gkentsidis, Mathieu Sarrazin, Sarah Barendswaard, Varun Kotian, and Riender Happee. 2025. Facial Emotion Recognition does not detect feeling unsafe in automated driving. arXiv:2509.04490 [cs.CV]. doi:10.48550/arXiv.2509.04490

[5] Xiaolin He, Jork Stapel, Meng Wang, and Riender Happee. 2022. Modelling perceived risk and trust in driving automation reacting to merging and braking vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour* 86, January (2022), 178–195. doi:10.1016/j.trf.2022.02.016

[6] Soyeon Kim, René van Egmond, and Riender Happee. 2024. How manoeuvre information via auditory (spatial and beep) and visual UI can enhance trust and acceptance in automated driving. *Transportation Research Part F: Traffic Psychology and Behaviour* 100 (Jan. 2024), 22–36. doi:10.1016/j.trf.2023.11.007

[7] Ruth Madigan, Tyron Louw, Marc Wilbrink, Anna Schieben, and Natasha Merat. 2017. What influences the decision to use automated public transport? Using UTAUT to understand public acceptance of automated road transport systems. *Transportation Research Part F: Traffic Psychology and Behaviour* 50 (2017), 55–64. doi:10.1016/j.trf.2017.07.007

[8] Abinay Reddy Naini, Zhaobo K. Zheng, Teruhisa Misu, and Kumar Akash. 2025. Self-Supervised Learning-Based Multimodal Prediction on Prosocial Behavior Intentions. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. ISSN: 2379-190X. doi:10.1109/ICASSP49660.2025.10890650

[9] Minxue Niu, Zhaobo K. Zheng, Kumar Akash, and Teruhisa Misu. 2024. Beyond Empirical Windowing: An Attention-Based Approach for Trust Prediction In Autonomous Vehicles. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5615–5619. ISSN: 2379-190X. doi:10.1109/ICASSP48485.2024.10446116

[10] Sina Nordhoff, Victor Malmsten, Bart van Arem, Peng Liu, and Riender Happee. 2021. A structural equation modeling approach for the acceptance of driverless automated shuttles based on constructs from the Unified Theory of Acceptance and Use of Technology and the Diffusion of Innovation Theory. *Transportation Research Part F: Traffic Psychology and Behaviour* 78 (2021), 58–73. TLDR: Investigating the attitudes and acceptance of automated shuttles in public transport among 340 individuals physically experiencing the automated shuttle 'Emily' from Easymile in a mixed traffic environment on the semi-public EUREF campus in Berlin found that compatibility with current travel was the strongest predictor of behavioural intention to use automatedShuttle.. doi:10.1016/j.trf.2021.01.001

[11] Daniela Paddeu, Graham Parkhurst, and Ian Shergold. 2020. Passenger comfort and trust on first-time use of a shared autonomous shuttle vehicle. *Transportation Research Part C* 115, March (2020), 102604. doi:10.1016/j.trc.2020.02.026

[12] Chen Peng, Pavlo Bazilinskyy, Yueteng Yu, and Natasha Merat. 2025. Measuring Passengers' Comfort and Perceived Safety in Automated Driving: Good Practices, Challenges, and Opportunities. In *Adjunct Proceedings of the 17th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, Brisbane QLD Australia, 298–301. doi:10.1145/3744335.3749142

[13] Chen Peng, Stefanie Horn, Ruth Madigan, Claus Marberger, John Lee, Josef Krems, Matthias Beggiato, Richard Romano, Chongfeng Wei, Ellie Wooldridge, Riender Happee, Marjan Hagenzieker, and Natasha Merat. [n. d.]. Conceptualising user comfort in automated driving: Findings from an expert group workshop. doi:10.31234/osf.io/76k8w

[14] Jork Stapel, Alexandre Gentner, and Riender Happee. 2022. On-road trust and perceived risk in Level 2 automation. *Transportation Research Part F: Traffic Psychology and Behaviour* 89 (Aug. 2022), 355–370. doi:10.1016/j.trf.2022.07.008

[15] Kathryn Wheeler. 2025. How to use Generative AI to Assist the Analysis of Qualitative Data. *Sage Research Methods How to Guides* (July 2025). https://repository.essex.ac.uk/40034/

Received 18 June 2026