LLMs for Explainable Automated Vehicles: Supporting Indonesia's Vision for 2045

Lucas Elbert Suryana
Department of Transport and Planning, 2628 CN Delft, the Netherlands
E-mail: l.e.suryana@tudelft.nl

ABSTRACT

Indonesia's vision for 2045, to be recognized as a developed country, is now assessed beyond income levels but also by innovation capacity, as reflected in the Global Innovation Index (GII), according to Kementerian PPN/Bappenas (2019). While Indonesia ranks 54th in the GII 2024, outperforming many peers in market sophistication and institutions, it still lags in R&D investment, human capital, and technology outputs (World Intellectual Property Organization, 2024). Achieving the 2045 goal; therefore, demands bold adoption of digital and technological innovations, together with stronger research capacity, that not only boost competitiveness but also fit the country's unique social and infrastructural realities. Among the frontier technologies that could narrow this innovation gap are automated vehicles (AVs), recognized for their potential to enhance safety and efficiency (Fagnant & Kockelman, 2015). In Indonesia, however, its relevance is shaped by unique challenges: motorcycle-dominated flows, informal transport, and weak rule enforcement. These conditions demand AVs that can navigate heterogeneous traffic using implicit social cues, unlike the structured settings of developed countries.

Although AV adoption is not yet imminent, past technological transitions, such as from horse-drawn vehicles to cars or landlines to mobile phones show that integration accelerates once adoption reaches critical mass. Proactive engagement is thus strategically significant, both to prepare AVs for Indonesia's traffic realities and to align with Vision 2045's innovation agenda. Early exploration of explainability and human-centered AV decision-making offers a pathway to regulatory readiness, public trust, and safer, socially aligned mobility.

Human drivers often rely on social cues and implicit reasoning when making choices in specific traffic situations (Zgonnikov et al., 2024). By contrast, current AI-based driving automation controllers are typically designed to follow strict optimization rules, which allow them to act consistently but provide little transparency about why certain decisions are made. The framework of meaningful human control (MHC) offers a potential way to address this gap by emphasizing the importance of aligning system behavior with human reasons, such as moral values and intentions (de Sio & den Hoven, 2018). However, existing AI models rarely make explicit whether their behavior reflects such reasons. Prior work has identified 12 categories of human reasons relevant for AV decision-making (Suryana et al., 2025), which could serve as a foundation for guiding AVs to better align their decisions with human reasons. Large Language Models (LLMs), with their ability to generate context-rich and human-like explanations, present a promising tool to operationalize this idea. Building on this potential, the present study asks whether LLMs can generate human-understandable explanations for AV decisions.

This research investigates whether Large Language Models (LLMs) can be leveraged to bridge this gap by generating human-understandable explanations for AV decision-making. In this study, LLMs are not tested as AV controllers. Instead, they are implemented as explanatory modules to existing controllers, designed to justify and clarify decisions made in challenging traffic scenarios. We use video-based traffic scenarios generated from a driving simulation system as the basis for prompting the LLM. These simulated situations capture ethically challenging conditions typical of Indonesian traffic, such as overtaking motorcycles or waiting at crowded intersections. Using OpenAI models as a case study, we ask the LLM to explain and justify actions within these scenarios. The prompts include the 12 categories of human reasons identified in prior work, and we evaluate the model's ability to integrate these reasons into its explanations. To validate the outputs, we compare the LLM-generated justifications with expert assessments of what an AV should do in the same situations. To further test

responsiveness to human reasons, we varied parameters in the prompts—for example, waiting time to reflect time efficiency and time-to-collision (TTC) to reflect safety.

The results indicate that the addition of human reasons increased alignment between LLM-generated decisions and those preferred by human experts. The models were able to recognize measurable variables, such as waiting time, and incorporate them into their reasoning, which enhanced the likelihood of safe and socially acceptable maneuvers. However, critical limitations remain. Generating robust explanations required an average of around 12 seconds, making them impractical for real-time, safety-critical driving contexts. Reliability was also inconsistent, with stable outputs often necessitating multi-run voting strategies, such as a two-out-of-three approach. Taken together, these findings underline that LLMs are unsuitable yet as primary AV controllers but promising as explanatory modules that can complement Advanced Driver Assistance Systems (ADAS), regulatory testing frameworks, and trust-building initiatives in Indonesia.

Explainable decision-making carries two important implications for Indonesia. First, it can support the development of regulatory frameworks that prioritize safety and transparency in emerging AV policies. Second, it can help foster public trust in digital technologies, a factor that is crucial for adoption in the diverse and informal traffic environments characteristic of Indonesian cities. By explicitly situating AV explainability within Indonesia's heterogeneous traffic conditions, this study contributes not only to international debates on AI safety but also to Indonesia's innovation strategy for Vision 2045. It highlights how human-centred explanations can bridge global AI advances with the social realities of Indonesian mobility.

Keywords: *Large Language Models, Explainable Decision-Making, Automated Vehicle.*

References

Fagnant, D. J., & Kockelman, K. (2015). Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. Transportation Research Part A: Policy and Practice, 77, 167-181.

Kementerian Perencanaan Pembangunan Nasional/Badan Perencanaan Pembangunan Nasional. (2019). Visi Indonesia 2045: Ringkasan eksekutif. Kementerian PPN/Bappenas.

Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. Frontiers in Robotics and AI, 5, 323836.

Suryana, L. E., Calvert, S., Zgonnikov, A., & van Arem, B. (2025). Principles and Reasons Behind Automated Vehicle Decisions in Ethically Ambiguous Everyday Scenarios. arXiv preprint arXiv:2507.13837.

World Intellectual Property Organization. (2024). Global Innovation Index 2024: Unlocking the promise of social entrepreneurship (17th ed.). WIPO.

Zgonnikov, A., Beckers, N., George, A., Abbink, D., & Jonker, C. (2024). Nudging human drivers via implicit communication by automated vehicles: Empirical evidence and computational cognitive modeling. International Journal of Human-Computer Studies, 185, 103224.