Human Driver's Reasoning on Moral Dilemma of Autonomous Vehicles: Values and Themes

Gi-bbeum Lee[1], Namwoo Kang[2], Ji-Hyun Lee[1]

[1] Graduate School of Culture Technology, Korea Advanced Institute of Science and Technology, South Korea

[2] The Cho Chun Shik Graduate School of Green Transportation, Korea Advanced Institute of Science and Technology, South Korea

ABSTRACT

The computing capacity of Autonomous Vehicles (AVs) has allowed the public to rediscover the classic trolley dilemma in a modern context. This paper aims to present an in-depth explanation of driver's moral reasoning for AV moral dilemma situations. Moral dilemma vignettes for AVs were designed based on real crash data and in-depth interviews with drivers. With the vignettes, a thought experiment with 33 participants was conducted; think aloud method and open-ended interviews were used to examine participants' reasoning processes. This paper qualitatively interpreted the data by thematic analysis. The results suggest that 1) human drivers' moral reasoning relies on diverse moral values and 2) reasoning can be explained based on safety, justice, and crash context. The results can be used as an analysis and communication tool for AV engineers and machine ethicists to determine how well current AV algorithms convey actual human moral reasoning.

Keywords: Autonomous vehicle, moral reasoning, moral dilemma, road transport, transportation, thematic analysis

INTRODUCTION

The computing capacity of Autonomous Vehicles (AVs) has allowed the public to rediscover the classic trolley dilemma in a modern context. AVs can make optimized decisions with advanced sensors, algorithms, and control, even during accidents. Consequently, the moral dilemma of AVs has become a question as the industry prepares for wide use of AVs. While the public anticipates that AVs will make utilitarian decisions to save more people, relying on a specific ethical principle can overlook the complexity of moral dilemmas (Awad *et al.*, 2018; Bonnefon *et al.*, 2016; Gordon, 2020; Malle *et al.*, 2015). There is a discrepancy between utilitarian choices in hypothetical dilemmas and what people would actually choose in reality (Bostyn *et al.*, 2018; Grasso *et al.*, 2020; Kallioinen *et al.*, 2019). Specifically, potential consumers prefer to buy the AVs that prioritize their passengers and are reluctant to agree on a policy for utilitarian vehicles (Bonnefon *et al.*, 2016). Viewpoints on the situation also affect moral judgments and moral confidence (Kallioinen *et al.*, 2019).

This paper extends the above studies to show that consequentialism alone will not be socially acceptable in the AV moral dilemma. We focus on the process through which humans choose ethical values and themes by asking three questions. First, "when can AVs face a moral dilemma in real-life traffic accident situations?" Second, "what factors can be considered in AV moral dilemmas from the driver's perspective?" Third, "what role do these factors play in driver's moral reasoning, and how can we understand and explain the reasoning process?"

While answering the questions, we aim to present an in-depth explanation of driver's moral reasoning for AV moral dilemma situations. Moral dilemma vignettes for AVs were designed based on real crash data and in-depth interviews with drivers. Using the vignettes, a thought experiment was conducted using the think aloud method and open-ended interviews to capture participants' reasoning processes. The data were analyzed using thematic analysis leveraging graph representation. We then discussed the themes of drivers' moral reasoning.

RELATED WORKS

Recent studies have investigated over a million sets of human choice data on AV moral dilemmas (Awad *et al.*, 2018; Bonnefon *et al.*, 2016; Frank *et al.*, 2019). The results shared a general inclination for saving a more significant number of people. Experimental studies using simulation methods such as virtual reality have extended the interpretation of moral decision-making to AV dilemmas (Grasso *et al.*, 2020; Kallioinen *et al.*, 2019; Li *et al.*, 2019). For example, Kallioinen *et al.* (2019) discovered that different perspectives influence moral judgments, while many crash studies were restricted to a third-party viewpoint. The researchers also showed that the detached observer perspective lessened self-confidence in the moral decisions. Meanwhile, researchers used hybrid approaches to discover contexts and reasons for

human judgments about moral dilemmas involving AVs. Qualitative insight from human moral decision-making can provide a background for explainability of artificial judgments (Arrieta *et al.*, 2020). If intuitive decisions are used to train a model for moral decisions without explainability, the system would cause an unintended, shifted evaluation of moral responsibility (Danielson, 2015).

The present work extends the above studies that investigated multiple perspectives of moral decision-making in AV cases. Although recent studies have provided quantitative descriptions of general tendencies, in-depth explanations of human morality in emerging cases have been lacking. Therefore, we start with qualitative methods to investigate moral reasoning in personal dilemma scenarios.

METHODS

This study consists of an AV moral dilemma vignette design, thought experiment, and thematic analysis. Details of our vignettes and thought experiment were introduced in previous works (Lee, 2018; Lee *et al.*, 2020).

Vignette Design

We investigated situations in which AVs could not avoid crashes based on the Special Crash Investigation (SCI) database of The National Highway Traffic and Safety Administration (NHTSA). Reference cases were collected from the database as follows: 1) unexpected behavior of another vehicle caused the crash, 2) other vehicles or obstacles were around the vehicle, and 3) crash caused at least severe injuries.

In-depth interviews were conducted to develop the above cases into moral dilemma vignettes. We recruited six adult drivers through convenience sampling. The reference cases were presented, and the subjects described any moral conflicts they experienced in the situations. Then, concepts of the moral conflicts were organized using the affinity diagram method. Based on the concepts, we developed moral dilemma vignettes that are applicable for AV crashes.

Thought Experiment

Adults of various ages, jobs, and levels of driving experience were recruited through social network services and online communities. Participants were paid around USD 10 for their participation. A total of 33 Korean adults participated in the experiment.

The experiments were conducted face-to-face and via videoconference while showing the vignette diagrams. Think aloud method and open-ended interviews were used as data gathering tools. Participants were asked to imagine themselves as drivers of the target vehicles in the vignettes, about to face crashes. They verbalized their every thought and emotion while reading the vignette and making their decision. Open-ended interviews followed to clarify language choices and capture participant intent.

Verbal recordings were transcribed as think aloud protocol. The researchers performed the data familiarization process and generated initial codes by line-by-line coding. The researchers labeled isolated sentences according to topics until the protocols of every participant had been coded. The codes were elaborated and structured in hierarchical categories.

Thematic Analysis

The protocols were converted to graphs consisting of nodes and edges. We used codes and code relations as nodes and edges in graphs. Based on the graphs, thematic networks were developed to visualize themes of human driver moral reasoning in crash situations.

Relations between codes in the decision process were found by carefully examining the protocols. Two relations, positive and negative, were defined between the codes (Table 1). If two codes had a positive relation, one code supported the impact of the other code. On the contrary, one code opposed the other when the codes had a negative relation, weakening the impact of the other code.

Table 1: Types of code relations

Relation type	**Meaning**
Positive relation	A supports B. If A is strengthened, B is strengthened.
Negative relation	A opposes B. If A is strengthened, B is weakened.

We identified themes that provided meaningful descriptions for our third research question - "what role do these factors play in driver's moral reasoning, and how can we understand and explain the reasoning process?" The themes consisted of sub-themes that provided detailed descriptions. First, to identify a sub-theme, we singled out a code that was frequently related to another code and put them on an empty graph as an initial state. Second, we added another code related to the context of the current graph state as an adjacent node and edge. Finally, we repeated the second step until the graph showed a latent meaning underneath the encoded data. After sub-themes were developed, they were placed together to form a thematic map of individual themes (Figure 1).

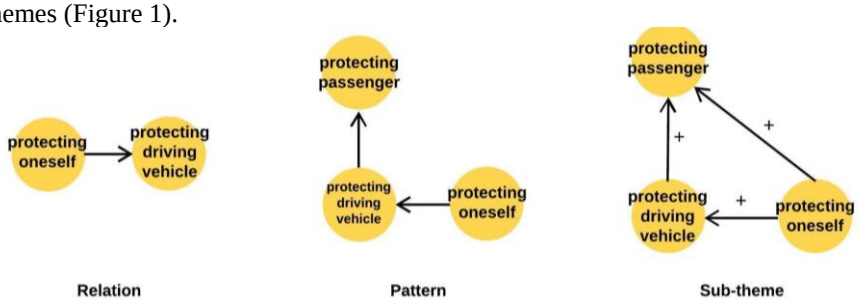

Figure 1: Developing thematic maps of driver's moral reasoning. Plus ("+") sign indicates a

positive relation between two codes.

RESULTS

This section introduces two results: driver value codes extracted from protocols and driver reasoning themes developed from thematic maps.

Driver Value Codes

The drivers' moral codes for the AVs moral dilemma were identified and grouped into three categories (Table 2). The categories were consistent with the concepts of the decision-making process, including normative aspects, procedural aspects, and actions (Rowe, n.d.). Cohen's kappa coefficient was calculated to check interrater agreement; the resulting value (K=0.35) indicated a *fair agreement* between the two raters. The codes and their frequency were introduced in (Lee *et al.*, 2020).

Driver Moral Reasoning Themes

Three themes of drivers' moral reasoning were extracted: safety-oriented reasoning, justice-oriented reasoning, and situational reasoning.

Safety-Oriented Reasoning. The theme is concerned with the safety of the driving vehicle, including occupants and the driver (Fig. 2). This theme was discovered in the data of 14 participants (42.4%). They identified the safety of themselves and their passengers and showed emphatic responses when they were at risk of serious injury. They often used terms of *empathy* for themselves and a sense of *driver responsibility*. The following are representative quotations: *"I think it is my responsibility to minimize the harm of an accidents for my vehicle." (P2) "If four children (passengers) are to die, there is a high probability that I will die as well." (P13)*

Justice-Oriented Reasoning. This theme is mainly based on two normative codes: minimization *of casualties* and *responsibility for fault* (Fig. 2). This theme was discovered in the data of 10 participants (30.3%). They focused on the lives of others rather than on their own lives or others' serious injuries. Although the participants assumed that a negligent driver would have to suffer the most significant damage, they made it the highest priority to *minimize casualties* rather than to take *responsibility for fault*. The following are representative quotations: *"Serious injuries can be recovered from by going to the hospital. So, I will choose the option with the lowest probability of death." (P29) "I think he made a severe mistake that put his life at risk. But if I could exchange the death of another person with my injury, I would do so." (P16)*

Situational Reasoning. This theme centered on the moral emotion of *guilt* (Fig. 2). This theme was discovered from the data of 9 participants (27.3%). The guilt-centered graphs generally showed a complicated, diverse range of codes (normative ideas, procedural ideas, and actions). The following are representative quotations: *"If*

someone dies, how am I going to survive it. How much would the passenger's family blame me?" (P17) "It may sound like I am egocentric, but I do not want to hit and kill a person. That would be very difficult. I don't think I can mentally handle such a situation." (P15)

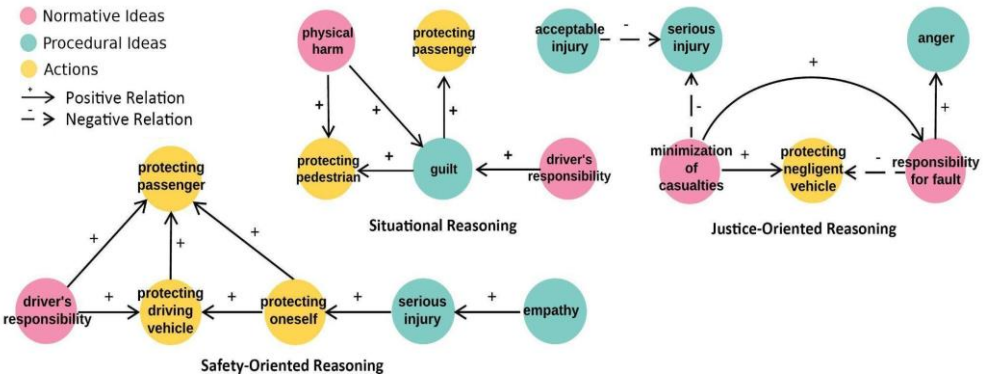

Figure 2: Themes of moral reasoning in graph representation.

DISCUSSION

The variety of normative ideas (what you should do) and actions (whom you should preserve) shows that a broad spectrum of prescriptive morality is applicable to the AV moral dilemma. The diverse codes show the diversity of moral values considered by drivers during moral decision-making. Even if two people choose identical crash consequences, how and why they made such decisions are affected by different values. Also, even if people perceive the same norms, other varying values can lead them to different choices. Therefore, for an AV to play the role of a moral agent, it should consider a broad spectrum of practical values rather than relying on a single moral theory.

The analysis in this paper confirms that drivers are oriented to safety, justice, and context when making moral reasoning in crash scenarios. Safety-oriented reasoning and justice-oriented reasoning are specifically contrasted. We found that justice-oriented reasoning aligned with previous findings of tendencies of utilitarian, law-sensitive judgments in AV moral dilemma situations (Awad *et al.*, 2018; Frank *et al.*, 2019; Li *et al.*, 2019). Safety-oriented reasoning also provided a qualitative explanation of collectivist moral decisions, as was also found in Awad et al. (2018). On the other hand, safety-oriented reasoning provides supporting evidence for the qualitative insights of Danielson (2015), in which participants shifted blame from robot drivers to humans. Safety-oriented reasoning showed the context in which human drivers take moral responsibility for their choices. In contrast, robot drivers imply an absence of direct responsibility, or what is called the responsibility gap

(Soltanzadeh *et al.*, 2020), thereby causing a perverse effect of shifting blame to others.

The results provide evidence to support the findings of utilitarianism and deontology in the area of AV ethics. Citizens in Korean culture have an awareness of utilitarian doctrine, as with the global trend (Awad *et al.*, 2018). The codes of *fault responsibility* and *protecting pedestrians* were aligned with the perspective effect (Frank *et al.*, 2019). This is also consistent with the finding that, in the context of moral dilemmas, Chinese drivers highly depend on whether there is a legal fault (Li *et al.*, 2019). The themes of safety, justice, and context were analogous to the findings of Bergmann et al. (2018), in which German participants made moral decisions from the driver's perspective. Although the group names differed from those in our research (Moral Egoists, Altruists, Switchers, and Unidentifiable), the descriptions for each group are similar to our three themes. The consistent explanations of human driver reasoning added reliability to this case study. Also, the cardinal concern with safety aligns with previous findings (Macioszek and Kurek 2020), in which the authors confirmed traffic users' subjective safety. Moreover, our qualitative data provide in-depth descriptions that support the quantitative results of Bergmann et al. (2018). Given the difficulty in making inferences about internal processes from quantitative data, our results help researchers obtain rich ideas about driver moral decisions.

Our moral values are focused on human morality; these values should not be directly conveyed to robot drivers (AVs). Typically, the evaluation of *protecting oneself* should be adjusted to align with human dignity. Nevertheless, our descriptive data on moral reasoning revealed the widespread permeation among drivers of the code of *protecting passengers*. *Driver's responsibility* has not been explicitly found in the quantitative studies (Bergmann *et al.*, 2018; Li *et al.*, 2019). This shows human drivers' strong belief that passenger safety is the driver's responsibility. Hence, we underline that *protecting passengers* is not only a market expectation but also within the scope of driver morality.

CONCLUSION

We leveraged the qualitative approach to propose themes of moral reasoning with detailed moral codes in cases of AV moral dilemmas. *Safety-Oriented Reasoning* was derived from driver responsibility and empathy for oneself and was faithful to the idea of the safety of oneself and passengers. *Justice-Oriented Reasoning* focused on norms such as minimization of casualties (a utilitarian value) and responsibility for fault. *Situational Reasoning* led to different decisions depending on contexts, showing sub-themes of the other two themes and the great impact of guilt. Various codes and themes show the diversity of human drivers' morals in unavoidable vehicle crash situations. Although our experiment was restricted to Korean, the results can be extended to make cultural comparisons by providing the case of an Eastern, collectivist culture.

The results provide qualitative insight for integrating human morality and intelligent systems. Our graph representation can be used as a qualitative tool to analyze AV judgment algorithms in emergencies by comparing those algorithms with human driver moral reasoning. Given the difficulty of cooperation of ethicists and engineers in manufacturing AVs, our results can serve as a communication tool to develop algorithms and convey moral perspectives of drivers, who are the potential users of AVs. The opening of such a discussion among ethicists and engineers, or other stakeholders, including potential users, will benefit social consensus for AV ethics. The results are also expected to help engineers think about differences between AV and human judgments in moral dilemma situations and allow them to better tune algorithms. AV engineers can use our vignettes as representative scenarios of moral dilemmas in real traffic situations.

ACKNOWLEDGMENTS

This work was supported by the BK21 plus program through the National Research Foundation (NRF), funded by the Ministry of Education of Korea.

REFERENCES

Arrieta, A.B., Díaz-Rodríguez, N., Ser, J.D., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., *et al.* (2020), "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI", *Information Fusion*, Vol. 58, pp. 82–115.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., *et al.* (2018), "The Moral Machine experiment", *Nature*, Vol. 563 No. 7729, pp. 59–64.

Bergmann, L.T., Schlicht, L., Meixner, C., König, P., Pipa, G., Boshammer, S. and Stephan, A. (2018), "Autonomous Vehicles Require Socio-Political Acceptance—An Empirical and Philosophical Perspective on the Problem of Moral Decision Making", *Frontiers in Behavioral Neuroscience*, Vol. 12, p. 31.

Bonnefon, J.-F., Shariff, A. and Rahwan, I. (2016), "The social dilemma of autonomous vehicles", *Science*, Vol. 352 No. 6293, pp. 1573–1576.

Bostyn, D.H., Sevenhant, S. and Roets, A. (2018), "Of Mice, Men, and Trolleys: Hypothetical Judgment Versus Real-Life Behavior in Trolley-Style Moral Dilemmas", *Psychological Science*, Vol. 29 No. 7, pp. 1084–1093.

Danielson, P. (2015), "Surprising judgments about robot drivers: Experiments on rising expectations and blaming humans", *Etikk i Praksis - Nordic Journal of Applied Ethics*, Vol. 9 No. 1, pp. 73–86.

Frank, D.-A., Chrysochou, P., Mitkidis, P. and Ariely, D. (2019), "Human decision-making biases in the moral dilemmas of autonomous vehicles", *Scientific Reports*, Vol. 9 No. 1, p. 13080.

Gordon, J.S. (2020), "Building Moral Robots: Ethical Pitfalls and Challenges", *Science and Engineering Ethics*, Vol. 26 No. 1, pp. 141–157.

Grasso, G.M., Lucifora, C., Perconti, P. and Plebe, A. (2020), "Integrating Human Acceptable Morality in Autonomous Vehicles", in Ahram, T., Karwowski, W., Vergnano, A., Leali, F. and Taiar, R. (Eds.), *Proceedings of the 3rd International Conference on Intelligent Human Systems Integration (IHSI 2020): Integrating People and Intelligent Systems*, Vol. 1131, presented at the Intelligent Human Systems Integration 2020, Springer, Cham, Modena, Italy, pp. 41–45.

Kallioinen, N., Pershina, M., Zeiser, J., Nezami, F.N., Pipa, G., Stephan, A. and König, P. (2019), "Moral Judgements on the Actions of Self-Driving Cars and Human Drivers in Dilemma Situations From Different Perspectives", *Frontiers in Psychology*, Vol. 10, p. 2415.

Lee, G. (2018), *The Analysis of Human Moral Reasoning for The Moral Decision-Making of Autonomous Vehicles (in Korean)*.

Lee, G., Rhim, J., Kang, N. and Lee, J.-H. (2020), "Driver Moral Codes in Autonomous Vehicles Dilemma Scenarios from Human Driver's Perspective (in Korean)", *Design Research*, Vol. 5 No. 1.

Li, S., Zhang, J., Li, P., Wang, Y. and Wang, Q. (2019), "Influencing factors of driving decision-making under the moral dilemma", *IEEE Access*, Vol. 7, pp. 104132–104142.

Macioszek, E. and Kurek, A. (2020), "Roundabout users subjective safety - case study from Upper Silesian and Masovian Voivodeships (Poland)", *Transactions on Transport Sciences*, Vol. 11 No. 2, pp. 39–50.

Malle, B.F., Scheutz, M., Arnold, T., Voiklis, J. and Cusimano, C. (2015), "Sacrifice One For the Good of Many? People apply different moral norms to human and robot agents", *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction*, presented at the The 10th Annual ACM/IEEE International Conference on Human-Robot Interaction, IEEE, pp. 117–124.

Rowe, P.G. (n.d.), *Design Thinking*, MIT press.

Soltanzadeh, S., Galliott, J. and Jevglevskaja, N. (2020), "Customizable Ethics Settings for Building Resilience and Narrowing the Responsibility Gap: Case Studies in the Socio-Ethical Engineering of Autonomous Systems", *Science and Engineering Ethics*, Vol. 26 No. 5, pp. 2693–2708.