A five-step ethical decision-making model for self-driving vehicles: Which (ethical) theories could guide the process and what values need further investigation?

Franziska Poszler[(1)]✉, Maximilian Geisslinger[(2)], Christoph Lütge[(1)]

(1) Peter Löscher Chair of Business Ethics; Institute for Ethics in Artificial Intelligence, Technical University of Munich (TUM), Arcisstraße 21, 80333 München, Germany
(2) Institute of Automotive Technology, Technical University of Munich (TUM), Boltzmannstr. 15, 85748 Garching b. München, Germany

✉ franziska.poszler@tum.de

Abstract

By choosing a specific trajectory (especially in accident situations), self-driving vehicles (SDVs) will implicitly distribute risks among traffic participants and induce the determination of traffic victims. Acknowledging the normative significance of SDVs' programming, policymakers and scholars have conceptualized what constitutes ethical decision-making for SDVs. Based on these insights and requirements formulated in contemporary literature and policy drafts, this article proposes a five-step ethical decision model for SDVs during hazardous situations. In particular, this model states a clear sequence of steps, indicates the guiding (ethical) theories that inform each step, and points out a list of values that need further investigation. This model, although not exhaustive and resolute, aims to contribute to the scholarly debate on computational ethics (especially in the field of autonomous driving) and serves practitioners in the automotive sector by providing a decision-making process for SDVs during hazard situations that approximates compliance with ethical theories, shared principles and policymakers' demands. In the future, assessing the actual impact, effectiveness and admissibility of implementing the here sketched theories, values and process requires an empirical evaluation and testing of the overall decision-making model.

Keywords: self-driving vehicle, autonomous driving, ethical decision-making, risk distributions

Introduction

Self-driving vehicles (SDVs) are one of the first commercialized AI-enabled robots to make decisions without human intervention, which will need to be pre-programmed (Liu & Liu, 2021). These decisions will carry ethical dimensions and are of normative significance (Dietrich & Weisswange, 2019) since SDVs will end up in situations with fatal consequences so that the programming of SDVs has palpable effects on road users in terms of traffic victims (Mordue et al., 2020). Given the risks involved and rapid technological developments, the investigation of how to program SDVs is a pressing matter (Nyholm & Smids, 2016). Therefore, policymakers have recognized the importance of considering ethical dimensions in programming SDVs (e.g., European Commission, 2021; U.S. Department of Transportation, 2016). Similarly, many scholars have engaged in discussing what should feed into the programming and what constitutes ethical decision-making for SDVs. For example, to approach this conceptualization, Poszler et al. (2023) have conducted a holistic review of the autonomous driving ethics literature, in which they evaluated the applicability of certain ethical theories and identified additional considerations (such as situation-adjusted risk distribution) that may prove useful to guide SDVs' ethical decision-making. Derived from an elaborated theoretical baseline including recent requirements formulated in policy drafts, this article proposes an explicit, compliant five-step ethical decision model for SDVs. To do so, this paper will be

structured as follows. First, theoretical fundamentals will be drawn from policymakers and contemporary scholars[1]. Second, the proposed model for ethical decision-making of SDVs will be highlighted by elaborating its decision process with an exemplary traffic scenario. Third, benefits of this model will be illustrated and underlying values that need concretization in the future will be pointed out[1]. Lastly, a short conclusion will be drawn by additionally emphasizing some caveats. Overall, although not exhaustive and resolute, this article aims to serve the scholarly community by contributing to the debate on computational ethics (especially in the field of autonomous driving) and practitioners in the automotive sector by laying out a potential solution for the 'ethical' programming of SDVs. In particular, this model approximates a decision-making process for SDVs during hazard situations that is compliant with ethical theories, shared principles and policymakers' demands and provides a checklist of underlying values that need further investigation in the future.

A glimpse into the results: A proposed model for ethical decision-making of SDVs

Based on the theoretical groundwork and identified (regulatory) requirements, this paper proposes a five-step model for ethical decision-making of SDVs. The sequence of steps is sketched in the following.

Step 1: Determination & calculation of possible trajectories. Decisions of SDVs are implemented via trajectory planning and selection. Thus, in a first step, the SDV needs to determine all potential trajectories and calculate corresponding consequences. Consequences that play a role in road traffic include, for example, passengers' comfort (determined by the vehicle's acceleration and jerk) and safety (i.e., physical integrity of the road users, determined by the risk posed to them).

Step 2: Typification of situation. The SDV determines the nature of the situation based on its ability to fulfill particular key duties. As the prime requirement for SDVs is safety, key duties (that are to be prioritized over values such as comfort) entail safeguarding the physical integrity of all traffic participants. The assessment of to what extent the SDV can comply with these rules/duties is determined in consultation with the risk values. If the SDV comes to the conclusion that at least one of the established rules/duties will be disobeyed (determined by the surpassing of a particular risk value), the SDV will declare a 'hazard situation mode' implicating a specific decision-making process (that is different to the 'non-hazard situation mode'). Namely, in the 'hazard situation mode', the only consequence to contemplate are the risk values, while other consequences such as the passenger's comfort or mobility are to be neglected.

Step 3: Exclusion of prohibited trajectories. To identify prohibited trajectories, the SDV checks the consequences of all trajectory alternatives for every traffic participant against particular risk thresholds. Values to be contemplated are overall risk as well as collision probability and estimated harm, each separately. If collision probability exceeds a particular threshold for at least one traffic participant, the value for estimated harm must not exceed a particular threshold and vice versa. Additionally, overall risk must not exceed a particular threshold. Those trajectories that fail to fulfill particular threshold restrictions are to be excluded; all remaining trajectory alternatives are reevaluated by the algorithm of the SDV in step 4.

Step 4: Calculation of valence-adjusted consequences. The SDV reevaluates all remaining trajectory alternatives by adjusting risk values with valence factors for different traffic participants. For example, traffic participants could be classified into pedestrians, cyclists and vehicles, with gradually declining valence factors. The new valence-adjusted risk values (vr) feed into the decision-making process of step 5.

Step 5: Selection of final trajectory. Based on the valence-adjusted consequences, the SDV calculates risk inequality (E) between all traffic participants as well as aggregated risk (U) for all trajectory alternatives. The aim is to identify the one trajectory that meets two distribution principles, namely: the greatest equal risk between traffic participants and that optimizes (i.e., minimizes) aggregated risk. To what extent E and U are factored in is pre-determined with a weighting factor for each consideration (we and wu). Given these weightings, the SDV can select the final action, i.e., the trajectory that has the lowest weighted-inequality-aggregated-risk value (W-E-U).

[1] These sections are not included in this extended abstract due to the word limit.

Conclusion

Overall, this paper aims to establish an ethical decision-making process for SDVs. In particular, this proposed model states where exactly which theories could hold during SDVs' decision-making and how additional considerations (i.e., concrete values) can be added in the calculation. Although not exhaustive and resolute, this approach may allow a first step towards correspondence with previously identified requirements from policymakers and scholars. Namely, the model utilizes overall risk (i.e., safety) as a central factor and takes into account the context, reasonableness and responsibility considerations as well as the protection of vulnerable road users. Furthermore, the model allows the integration of a mix of ethical theories and shared principles and overall, provides a chronological order for particular decision-making steps, while leaving room for future adjustments. Nevertheless, this proposed model is not without limitations and we need to be precautious that this proposed decision-making model for SDVs indeed derives from ethically grounded and 'justified' requirements and serves the benefit of society. To assess the actual impact, effectiveness and admissibly of implementing the here sketched theories, values and process, amongst other, a necessary next step is the empirical assessment[2] of this SDV decision-making process

References

Dietrich, M., & Weisswange, T. H. (2019). Distributive justice as an ethical principle for autonomous vehicle behavior beyond hazard scenarios. *Ethics and Information Technology*, 21(3), 227-239. https://doi.org/10.1007/s10676-019-09504-3

European Commission (2021). *Regulatory framework proposal on artificial intelligence*. Retrieved from: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Geisslinger, M., Poszler, F., & Lienkamp, M. (2023). An ethical trajectory planning algorithm for autonomous vehicles. Nature Machine Intelligence, 5(2), 137-144. https://doi.org/10.1038/s42256-022-00607-z

Liu, P., & Liu, J. (2021). Selfish or utilitarian automated vehicles? Deontological evaluation and public acceptance. *International Journal of Human–Computer Interaction*, 37(13), 1231-1242. https://doi.org/10.1080/10447318.2021.1876357

Mordue, G., Yeung, A., & Wu, F. (2020). The looming challenges of regulating high level autonomous vehicles. *Transportation research part A: policy and practice*, 132, 174-187. https://doi.org/10.1016/j.tra.2019.11.007

Nyholm, S., & Smids, J. (2016). The ethics of accident-algorithms for self-driving cars: An applied trolley problem?. *Ethical theory and Moral Practice*, 19(5), 1275-1289. https://doi.org/10.1007/s10677-016-9745-2

Poszler, F., Geisslinger, M., Betz, J., & Lütge, C. (2023). Applying ethical theories to the decision-making of self-driving vehicles: A systematic review and integration of the literature. Working Paper.

U.S. Department of Transportation & National Highway Traffic Safety Administration. (2016). *Federal Automated Vehicles Policy: Accelerating the Next Revolution in Roadway Safety*. Retrieved from: https://www.transportation.gov/AV/federal-automated-vehicles-policy-september-2016

[2] Geisslinger et al. (2023) can serve as a methodological example of such an empirical assessment, in which a simulation was conducted to show how risk distributions among traffic participants change when particular ethical theories are implemented into an SDV's trajectory planning algorithm.