Title:

Safety, Privacy, Fairness, Interpretability and Responsibility of Autonomous Driving

Authors:

AI4EU WG: Manuela Battaglini, Xin Chen, Ana Chubinidze, Francesca Foffano, Fabio Fossa, Teresa Scantamburlo, Zahoor Ul Islam, Ricardo Vinuesa.

Intro

This document presents preliminary work concerning the application of the European ethical guidelines on autonomous driving to an actual case study. It showcases the results of one of the research activities carried out by the AI4EU working group on Ethical, Legal, and Social Issues.

The purpose of our case study is two-fold. On the one hand, (a) we try to provide **concrete ethical recommendations** to the engineers involved in the project, with particular attention to what concerns the development of control logics and human-machine interfaces. On the other hand, (b) we wish to **test the effectiveness of the European framework** presented in the report *Ethics of Connected and Automated Vehicles* [ECAV] in supporting ethical analysis of autonomous-driving projects. This requires at least (i) determining which recommendations are relevant vis-a-vis the case study and which are not; (ii) specifying what relevant recommendations entail for the case study; and (iii) identifying further important aspects that recommendations fail to properly address. Even though (a) and (b) can be construed as two separate purposes, they imply each other, so they are simultaneously pursued.

After a brief overview of the case study, we discuss our findings following the ECAV tripartite structure, i.e., by focusing on safety (1), privacy-fairness-interpretability (2) and responsibility (3).

Case study

The case study on which we focus has been provided by engineers working at the Politecnico di Milano [POLIMI], Italy. The project focuses on the development of control logics and internal/external human-machine interfaces to tackle navigation problems in an **unstructured environment** with active interaction with **multiple vulnerable road users**. Such technologies are intended to enable a fully-automated electric shuttle to safely and effectively navigate through the Politecnico di Milano La Masa campus among its population of pedestrians, cyclists, e-scooter users, etc.. The project is part of a wider set of research activities dedicated to various aspects of autonomous driving and is still in its conceptualization stage. Consequently, the ethical analysis is aimed at providing guidance to the engineering team on how to properly address relevant ethical concerns since the very outset.

Safety

Safety concerns related to the case study are identified and discussed by considering ECAV recommendations 1–6. Even though to a varying extent, each recommendation applies to our case study. As a whole, they offer valuable guidance and succeed in covering potential safety issues. However, some challenges must be faced. Particular attention is to be dedicated to the establishment of meaningful **benchmarks** and **metrics** to assess safety in relation to crucial functions such as obstacle avoidance, collision avoidance, speed regulation, etc.. This implies selecting analogous scenarios for relevant comparison, which raises severe epistemological problems. Moreover, it also entails accounting for all relevant categories of road users, which might require in-depth empirical studies on campus population. Furthermore, the unstructured and shared nature of the environment where the shuttle will be tested and deployed poses many **safety risks** that must be appropriately identified, managed, and minimized. Also, given the pivotal role played by interfaces to ensure safety, a **design-for-all approach** that puts inclusiveness at the center must be pursued to assure that users can realize risks and act accordingly independently from disabilities or impairments. Finally, **emergency procedures** must be in place to safely handle malfunctioning, either autonomously operated or involving humans (which, however, poses its own challenges, e.g., misuses, overload, overreliance, etc.).

Privacy, fairness, interpretability

The essential aspects of **privacy, fairness and interpretability** in connection to this project are covered by the ECAV recommendations 7–15. We believe that, although these recommendations help to establish the general framework and context, they are limited in a number of aspects of critical importance. Firstly, there is not any explicit recommendation regarding **decentralized data gathering**, which should be a basic requirement for autonomous-vehicle applications. Although some of the gathered databases can be very valuable when it comes to research & development, their usage should be extremely controlled due to the numerous privacy and individual-right concerns associated with them. This may require a complete paradigm shift when it comes to data policy, where **consent is simply not enough**. In this sense, we believe that developing some indicator which explicitly marks when data is being gathered is essential. Then, the surrounding actors (including other cars, pedestrians, etc.) can be notified and proper data-gathering actions can be taken. Furthermore, when decisions are being made by the autonomous system (e.g. in terms of avoiding an accident), **complete transparency of the decision process** is required, together with the right to contest such decisions; in this application, accountability is very important. Transparency of autonomous vehicles can be achieved through the development of AI models which are interpretable from the beginning. The problem is the fact that higher performance in the required computer-vision tasks is obtained through deep learning, which does not allow interpretability of the model results. In this direction, we propose to use approaches based on **developing symbolic models** from the already-trained deep-learning models by means of inductive biases, a methodology that may help to increase the interpretability of these models.

Responsibility

The notion of responsibility is covered by ECAV's recommendations 16-20. Responsibility cannot straightforwardly be embedded into the product itself, so recommendations 16-20 rather seek to "foster a *culture of responsibility"* (ECAV, p. 53). Moreover, responsibility is distributed across **many stakeholders** (deployers, users, manufacturers, suppliers, regulators, market surveillance authorities, etc.), which makes it difficult to find unilateral actions for engineers to take. For example, most of the sub recommendations (7 out of 9) that concern "manufacturers and deployers" also require policymakers' collaboration (ECAV, pp. 68-69). The engineers working on the POLIMI project are therefore not the only target audience and will find themselves unable to fulfill the recommendations unless further guidance is issued or norms are agreed upon across stakeholders on how to share responsibility. A first step towards this will be to communicate to project managers/engineers the ECAV recommendations on responsibility and assist in raising **awareness** about them within the team – emphasizing the broader, multistakeholder effort underway to clarify the **distribution of responsibility**. A second step will be to gather engineers, manufacturers and deployers industry wide for them to converge on a way to discuss the questions of responsibilities raised by ECAV and how to comply with its recommendations, including defining the terms and time of a feasible implementation, and identifying the specific tools for industry-ready policies and practices.

Conclusion

To conclude, the ECAV recommendations offer **initial guidance** to discuss several ethical aspects of our case study. However, **further discussion** is needed both to meaningfully apply them and to identify other relevant issues.

For what concerns safety, **future work** must be dedicated to establishing reliable safety metrics and managing risks appropriately. In relation to privacy, fairness, and interpretability, a number of aspects of critical importance, such as decentralized data gathering and transparency, must still be properly highlighted. Finally, responsibility issues require the organization of concerted stakeholder action to be appropriately dealt with.