How to Program Fairness in Intelligent Systems: Breaking Status Quo Solutions by Using Explainable AI

Fernando Muñoz, Christian Servin, Nadia Karichev
Program of Computer Science, El Paso Community College, El Paso, TX, 79915

Abstract

With a possible shift in fuel resources, combined with recent technological advancements, it is becoming increasingly obvious that the motor automotive industry will begin to evolve and propose new ideas for everyday transportation. A concept that has gained more popularity in the past decade is the daily driving of Autonomous Vehicles (AVs). However, challenges emerge when these intelligent systems need to perform an ethical decision in a possible worst-case scenario. What happens when AVs are in a position such as a *trolley dilemma*. Current autonomous vehicles are modeled after traditional Artificial Intelligence (AI), which are known to encounter a 'black box problem', which refers to the issue concerning the unexplainable abrupt decisions due to poor modeling or lack of input while testing.

Introduction

Autonomous vehicles (AVs) are meant to increase the safety and comfort of humans, however, can these intelligent systems be trusted with the lives of humans, and can these intelligent systems perform better than a human in ethically challenging situations? In this research, we plan to recognize the notion of fairness (equity, equality, and justice) while designing intelligent systems, particularly, autonomous vehicles.

Background

Traditional AVs are modeled after traditional Artificial Intelligence (AI) and due to numerous possible situations that the AV might find itself in, the AI model will turn into a black-box model [4] and [6]. The use of Explainable Artificial Intelligence (XAI) will turn the traditional AI AV model into a safer, more socially acceptable, and fair intelligent system as demonstrated by **Figure 1.**

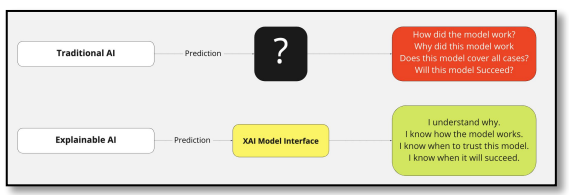

Figure 1. Differences between AI and XAI

Methodology

The proposed architecture for this research can be demonstrated by **Figures 2, 3 and 4.**

- **Figure 2** represents the danger levels that are encountered in the road every day.
- **Figure 3** represents the numerous amount of scenarios that may occur under each danger.
- **Figure 4** demonstrates the decision tree that would be used in order to decided the worst best-case outcome.

As part of an initiative in this research, individuals responded to a series of scenarios in a controlled survey using *Google Forms*. Each scenario as demonstrated by consist of a possible collision with an AV, **Figure 5**. Testees are to choose what they believe to be the best possible decision. Scenarios can be demonstrated by **Figures 6, 7, and 8.**

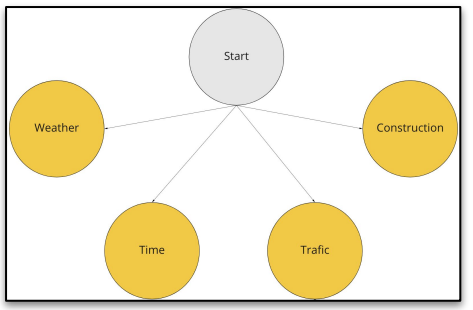

Figure 2. Danger Levels

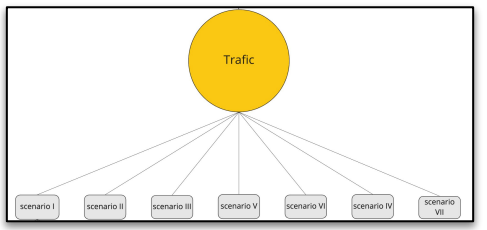

Figure 3. Scenarios

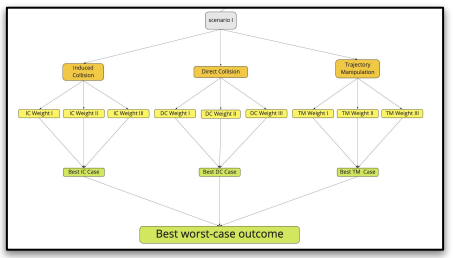

Figure 4. Proposed Decision Tree

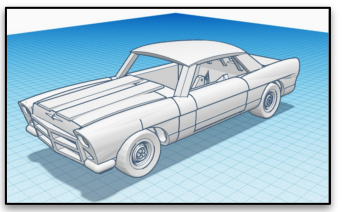

Figure 5. AV Model used for Scenarios

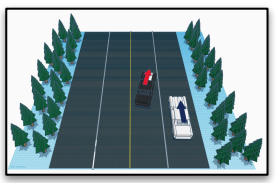

Figure 6. Low Level Danger Scenario

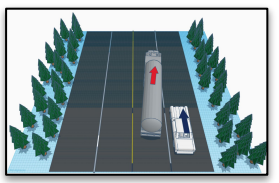

Figure 7. Medium Level Danger Scenario

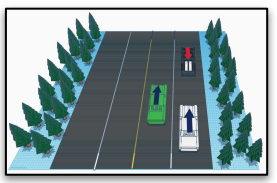

Figure 8. High Level Danger Scenario

Preliminary Results/Summary

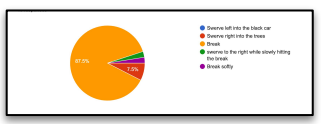

Figure 9. Low Level Danger Results

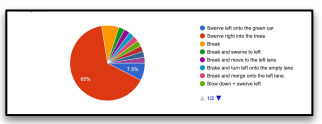

Figure 10. Medium Level Danger Results

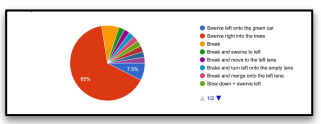

Figure 11. High Level Danger Results

In low level danger scenarios, the majority of testees chose the same decision, we can define this choice as the status quo. However, as more danger began to arise, testees began to derive from the common decisions and began to provide answers that differ from the status quo as seen in **Figures 9, 10, and 11**.

Future Work

- Scenarios will be migrated to a different platform, *Teacher Moments.*
- Credibility of the testes will be considered and determined when conducting experimentation.
- The results of these scenarios will be applied to a function to determine which weight promotes the status quo.
- Proposed Architecture for the learning model will be applied in a simulation to compare the effectivity between a traditional AI model and the proposed XAI model.

References

[1] C. Servin, V. Kreinovich, and S. Shahbazova, "Ethical Dilemma of Self-Driving Cars: Conservative Solution Ethical Dilemma of Self-Driving Cars: Conservative Solution," 2021J. Clerk Maxwell, *A Treatise on Electricity and Magnetism,* 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp. 68-73.
[2] C. W. Bauman, A. P. McGraw, D. M. Bartels, and C. Warren, "Revisiting external validity: concerns about trolley problems and other sacrificial dilemmas in moral psychology", *Social and Personality Psychology Compass*, 2014, Vol. 8, No. 9, pp. 536–554.
[3] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, 2019.
[4] D. Castelvecchi, "Can we open the black box of ai," *Nature*, vol. 538, no. 7623, p. 20, 2016.
[5] H. Mankodiya, M. S. Obaidat, R. Gupta and S. Tanwar, "XAI-AV: Explainable Artificial Intelligence for Trust Management in Autonomous Vehicles," *2021 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*, 2021, pp. 1-5, doi:
[6] P. Koopman and M. Wagner, "Autonomous vehicle safety: An interdisciplinary challenge," *IEEE Intelligent Transportation Systems Magazine*, vol. 9, no. 1, pp. 90–96, 2017.
[7] R. S. Basim Mahbooba, Mohan Timilsina and M. Serrano1,"Explainable artificial intelligence (xai) to enhance trust management in intrusion detection systems using decision tree model," 2021.
[8] V. Behzadan and A. Munir, "Adversarial Reinforcement Learning Framework for Benchmarking Collision Avoidance Mechanisms in Autonomous Vehicles," in *IEEE Intelligent Transportation Systems Magazine*, vol. 13, no. 2, pp. 236-241, Summer 2021, doi: 10.1109/MITS.2019.2898964.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 2034030 and 1834620. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.