Thesis Project Portfolio

Simulating Autonomous Vehicles for XAI

(Technical Report)

How Legal Systems Struggle to Maintain Accountability When AI Make Erroneous Decisions

(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Arran Scaife

Spring, 2023

Department of Computer Science

Table of Contents

<center>**Sociotechnical Synthesis**</center>

Artificial Intelligence (AI) has become a rapidly growing and transformative technology that has the potential to revolutionize many aspects of our lives from driving to writing. As such, it is crucial to ensure that AI is developed and used in an ethical and accountable manner. Thus, this synthesis includes two research projects connected by the principles and processes of building responsible and ethical AI. The first paper will focus on the development of a framework for explainable AI (XAI) using SafeBench, a software designed and used to seamlessly generate car simulation scenarios in a vehicle simulation software known as CARLA. The later aim of this research is to utilize this framework to detect out-of-distribution features from video of AI crashes that can be analyzed and used in legal proceedings to explain the decision-making process of the AI system. The second paper demonstrating the STS research will highlight the legal implications of AI systems that lack accountability, specifically under tort litigation, biased sentencing, and intellectual property, as well as the unintended consequences of relying on such systems. Together, both of these papers shed light on the importance of building ethical and accountable AI systems that can be trusted by providing means to protect individuals and prevent harm from AI's consequences.

The technical component of this project aims to investigate how SafeBench can be used to autonomously generate car simulation scenarios and later be used to detect out-of-distribution factors in AI images saved from car cameras. Towards this, the CARLA self-driving simulator helps to accurately simulate and generate crash data while SafeBench uses reinforcement learning and generative scenarios to train the CARLA agent. Throughout the paper, the agents' performance is visualized across several scenarios and figures and limitations and potential areas for future work are discussed afterwards. As AI becomes more ubiquitous, understanding their

decisions becomes increasingly crucial to hold actors legally accountable when unintended consequences inevitably arise. By identifying factors relevant to these decisions, such as out-of-distribution features, we can provide more accurate and persuasive explanations for AI decisions in court cases. Thus, this research provides a framework that stands as a crucial step towards making AI systems more transparent and explainable, making it possible to build safer and more trustworthy AI systems that can be used throughout a variety of contexts.

The STS component of this project investigates the many legal complications that arise as a result of AI systems that lack accountability. As dominating rise in society, especially in transportation, criminal justice, and art, indicates an increasing importance to understand AI's legal implications. This STS paper will explore how legal systems currently fail to determine accountable actors when AI make erroneous or unethical decisions that cause harm. By applying Susan Leigh Star's framework on infrastructure, the paper examines how traditional legal concepts such as fault, liability, and intellectual property, erode in the face of complex socio-technical systems involving AI decision-making.

Finally, both components of this project are designed to simultaneously address the overall socio-technical challenges of building ethnically accountable AI. By developing a framework for explainable AI while investigating the legal implications of AI decision-making, this project aims to contribute a more comprehensive understanding of AI's impact on societal functions. Therefore, integrating these technical and social perspectives offers a holistic approach to building AI systems that justifies a need for them to be transparent, accountable, and trustworthy so that we develop responsible AI designed to benefit society while minimizing the risks of unintended consequences associated with this powerful technology.