AI Ethics Field Guide Exhibit - Case & Mechanism Explanation - Final Version

This case is centered on a real-world safety and security failure involving an AI operated autonomous vehicle. In March 2018, a pedestrian, Elaine Herzberg, was struck and killed in Tempe, Arizona by an Uber self-driving test vehicle operating in autonomous mode. The vehicle was part of Uber's advanced driver-assistance and self-driving program, which relied on AI-based perception systems to detect objects, classify them (pedestrian, bicycle, vehicle), and decide whether braking or evasive action was needed. While a human safety driver was present, the system failed to correctly identify and classify Elain and the driver did not react in time to intervene and avoid the crash. This case shows how AI used in safety critical situations can cause severe negative outcomes to people, vehicles and overall road safety, when the technology, governance, and oversight don't meet the level of risk.

The key stakeholders in this case include Elain Herzberg and her family who experienced the most direct harm. Uber is a major stakeholder as they developed, operated, and are responsible for the autonomous system/vehicle used in the accident. The safety driver would be a stakeholder as they were responsible for monitoring and being aware of potential danger, as well as intervening when necessary. Some other stakeholders would include the engineers and managers of the vehicle's emergency detection and braking system. Arizona and the regulators who allowed this autonomous program to be in place and tests to occur on live roads. The final stakeholders would be normal pedestrians and civilians as they must share the road with these systems and face the risks of experiments failing.

The mechanism of harm occurred through many failures throughout the AI implementation pipeline. The model, workflow, and governance points are the clearest entrances for ethical concerns. At the model level the detection system failed to correctly identify and classify Mrs. Herberg as a pedestrian while she was walking her bike. This misclassification resulted in a delayed system response time and prevented timely braking. The workflow level is affected as the safety driver is entrusted to supervise a complex autonomous system alone and make split second adjustments in real-time. The governance failures played the most major role as Uber continued to carry out these tests while being aware of the heavy risks and limitations in safety. The result was a combination of AI system failure and human error in oversight that led to a fatal accident.