EVARA
Ethical Violation Analysis and Risk Assessment

May 26th 2024

Case Study on Human Welfare Violations Related to Bias in Self-Driving Cars.

AIVMDB Filtering Analysis Output (Sector: Automotive, All. Ethical Issue: Bias. Violated Value: Human Welfare)

Example Type	Title	Ethical Issue	Violated Value
Instance	Pedestrian Detection Dark Skin Bias 1	Bias - Skin Color	Universal Usability, Human Welfare
Instance	Pedestrian Detection Dark Skin Bias 2	Bias - Skin Color	Universal Usability, Human Welfare
Instance	Pedestrian Detection Age Bias (Children)	Bias - age	Universal Usability, Human Welfare
Instance	Tesla Obstacle Detection Failure (Tractor-Trailer)	Bias - Object detection/prediction	Human Welfare , Universal Usability
Scenario	Lack of training data collection diversity	Bias	Universal Usability, Human Welfare
Instance	Cruize pedestrian detection bias - children	Bias - age, Negligence	Human Welfare, Universal Usability, Accountability
Instance	Cruize vehicle collided with articulated vehicle due to failure to predict movement.	Bias - object detection/prediction	Universal Usability, Human Welfare
Scenario	Data injection attacks on pedestrian detection training data.	Compromised Security, Bias	Human Welfare, Universal Usability
Scenario	Data injection attacks targeting fairness metrics and injecting bias into the dataset	Compromised Security, Bias	Human Welfare , Universal Usability

System Definition:
Define the target of this evaluation or the system under development.
Self-driving car with level 4 or 5 autonomy

Asset/Function Identification:
Define the asset within the system being evaluated by EVARA.
Pedestrian Detection Model

MLLC Phase Identification:
Define the asset MLLC phase. (Data Management, Training, Testing, Deployment, full cycle)
Data Management (collection and labeling)

Value Identification and Value Violation Scenarios:
Identify the chosen value that is to be protected. Identify possible consequences or damage scenarios of compromise (Ecosystem, Organizational, Human)

Chosen Value: Human Welfare

Impact Rating:

List the different value violation scenarios and evaluate the possible impact in each of the 3 categories.
Ratings:(Severe, Major, Moderate, Negligible)

Value Violation		Ecosystem	Organizational	Human
VV1	Compromised training dataset	Negligible	Moderate	**Severe**
VV2	High miss rates or disparity in miss rates for pedestrians	Negligible	Major	**Severe**
VV3	false pedestrian trajectory prediction or disparity in pedestrian trajectory prediction.	Negligible	Moderate	**Severe**

Violation Scenario Identification and possible path analysis:
List each value violation scenario along with the possible paths for each scenario with respect to the chosen value, system asset, and MLLC phase.

VV1 - Violation Scenarios		possible path
VS1	Data poisoning adversarial attack	P1 - A label flipping attack done without proximity to the vehicle.
		P2 - A label flipping attack done with proximity to the vehicle.
		P3 - Data injection/poisoning attack done without proximity to the vehicle.
		P4 - Data injection/poisoning attack done with proximity to the vehicle.
VS2	Compromised data collection hardware	P1 - Compromised intake camera.
		P2 - Compromised intake lidar.
		P3 - Compromised intake radar.

VV2 - Violation Scenarios		possible path
VS1	Misrepresentation of populations in taring data	P1 - Data is collected from an area with low diversity.
		P2 - Data is collected from locations that are almost the same demographically, missing variations in pedestrian behavior and visual cues.
		P3 – training population does not align with deployment population
VS2	Improper data labeling practices	P1 - Labels for protected classes or minority populations are excluded from the training dataset.

		P2 - Labeling practices are not uniform or standardized leading to significant variation.

VV3 - Violation Scenarios		possible path
VS1	Misrepresentation of populations in training data.	P1 - Data is collected from an area with low diversity.
		P2 - Data is collected from locations that are almost the same demographically, missing variations in pedestrian behavior and visual cues.
		P3 – Training population does not align with deployment population
VS2	Improper data labeling practices	P1 - Labels for protected classes or minority populations are excluded from the training dataset.
		P2 - Labeling practices are not uniform or standardized leading to significant variation.
		P3 – Using automated labeling tools that's perpetuate biases or misconceptions.

Path Feasibility Ratings:
Rate each possible value violation path on likelihood. (Very High, High, Medium, Low, Very Low)
You may use the Likelihood scale recommendation tool to determine the rating.

Value Violation 1 (VV1)

VS1 - Violation path	Likelihood
P1	High
P2	High
P3	High
P4	High
VS2 - Violation path	Likelihood
P1	High
P2	High
P3	High

Value Violation 2 (VV2)

VS1 - Violation path	Likelihood
P1	Very High
P2	Very High
P3	Very High
VS2 - Violation path	Likelihood
P1	Very High

P2	Very High
Value Violation 3 (VV3)	
VS1 - Violation path	Likelihood
P1	Very High
P2	Very High
P3	Very High
VS2 - Violation path	Likelihood
P1	Very High
P2	Very High
P3	Very Low

Risk Scores

Based on the possible path feasibility level, and the value violation scenario impact rating, determine the risk level.

Use the highest reported level for each violation scenario (highest attack path feasibility and highest damage scenario impact rating)

Value Violation Scenario	Possible path feasibility (highest)	Damage scenario impact rating (highest)	Risk level
VV1-VS1	High	Severe	5
VV1-VS2	High	Severe	5
VV2-VS1	Very High	Severe	**6**
VV2-VS2	Very High	Severe	**6**
VV3-VS3	Very High	Severe	**6**
VV3-VS2	Very High	Severe	**6**

Impact\Likelihood	Very Low	Low	Medium	High	**Very High**
Severe	1	3	4	5	**6**
Major	1	2	3	4	5
Moderate	1	2	2	3	4
Negligent	1	1	1	1	3

Risk determination:

1- VV2
2- VV3
3- VV1