EVARA

Ethical Violation Analysis and Risk Assessment

The User Guide

Nada Madkour nmadkour@emich.edu
March 9th 2024
GitHub

Table of Contents

Nada Madkour nmadkour@emich.edu

March 9th 2024
GitHub

About EVARA

Purpose

The purpose of EVARA is to systematically evaluate the potential risks associated with activities, decisions, or technologies, focusing specifically on their potential to violate critical human values. Through comprehensive analysis and assessment, the tool aims to identify and prioritize ethical risks, enabling stakeholders to make informed decisions and mitigate potential harm to individuals, communities, and society as a whole. By integrating value-based considerations into risk assessment processes, we strive to promote responsible decision-making and uphold fundamental human values in various domains.

Applicability

EVARA is designed to be flexible and adaptable, making it applicable across various phases of the machine learning lifecycle. Whether in the early stages of system design, during development and training, or in the deployment and monitoring phases, our tool provides valuable insights to system developers, designers, and policymakers.

EVARA is designed to be user-friendly and accessible to a wide range of stakeholders, including system developers, designers, and policymakers with varying levels of technical expertise. Its modular and customizable nature allows users to tailor the assessment process to their specific context and requirements, ensuring relevance and effectiveness across diverse applications and domains within the machine learning lifecycle.

Pre-VARA

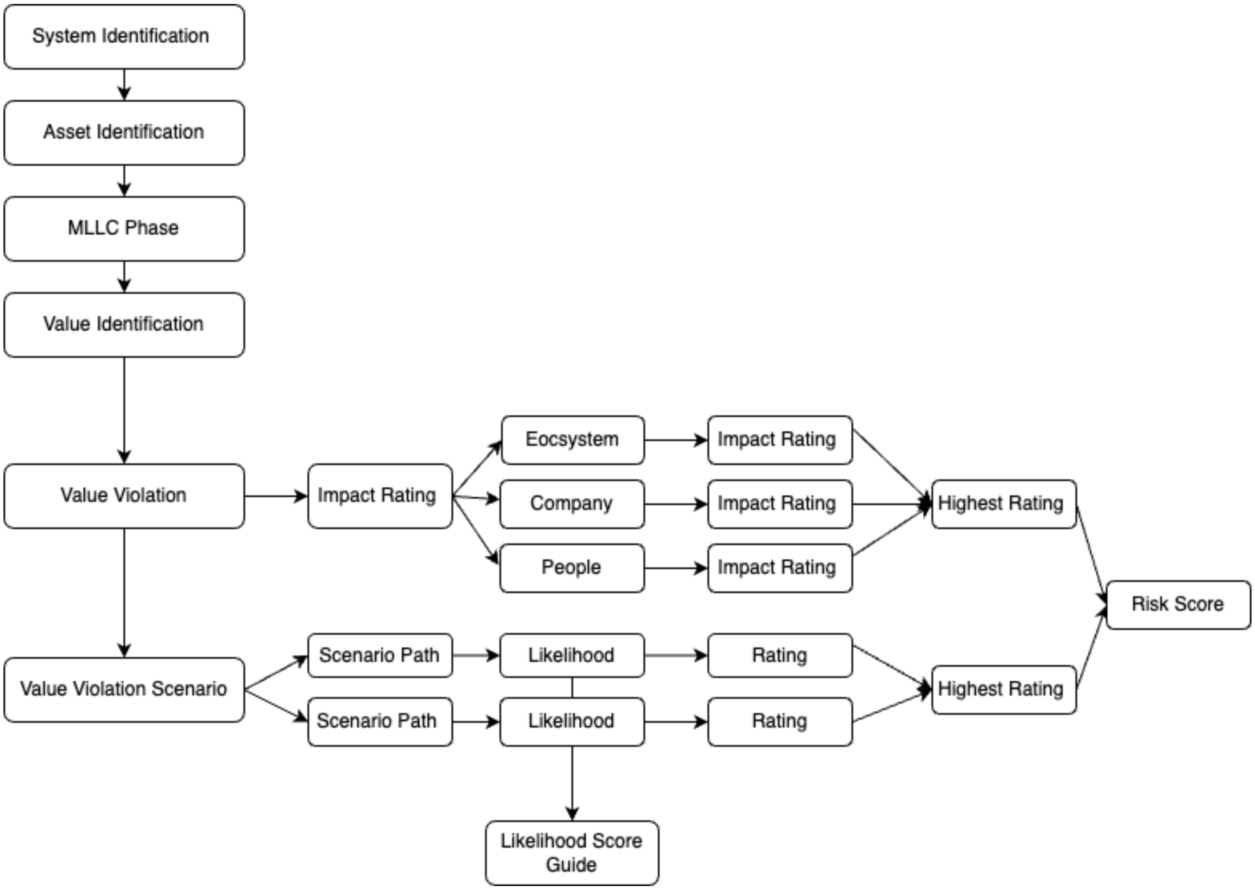
It is the responsibility of the user to conduct in-depth research on the value violation and the system being evaluated.

In preparation for using the tool, it is recommended that the following actions be considered:

1- Utilizing knowledge bases that report AI-related incidents.
   a. AIVMDB
   b. AIAAIC Repository
   c. Bias in AI: Example Tracker
   d. AI Incident Database
2- Gathering stakeholder perspectives. This can be done with interviews, surveys, or other methods.
3- Utilizing red teaming efforts.

Nada Madkour nmadkour@emich.edu
March 9th 2024
GitHub

E-VARA Structure



System Identification:

Identifying the system being evaluated by E-VARA. The system is meant to represent the overall item being evaluated.

Examples: Self-driving car, Melanoma diagnosis tool, Job application filtering tool.

Asset/Function Identification:

Identifying the asset or function within the system being evaluated by E-VARA. The asset of function is the specific portion of the system being evaluated. This portion can be a model, a function, an asset, or an algorithm.

Examples: Pedestrian detection model, Melanoma image recognition technology, Resume ranking algorithm.

MLLC Phase Identification: (optional)

Identifying the phase of the asset MLLC being evaluated. Identify the phase(s) of the machine learning lifecycle that are being included in the evaluation. This part is optional. You may choose one phase, multiple phases, or all phases.

Suggested machine learning lifecycle Phases:
Data management, Model training, Model testing, Deployment, Use and monitoring.
View more on the MLLC here.

Value Identification:

Identifying the value being assessed for risk. It is recommended that the AI Value Mapping Database (AIVMDB) be used for this step.

Value Violation:

List the value violations being assessed for risk. Each value violation should represent a high-level method that may result in value violation.
A value violation refers to any action, decision, or outcome, that results in the violation of the value. The value violation is meant to be a more high-level representation of the action/decision/outcome. The value violation paths will be used to represent the detailed steps or requirements of violation materialization.
Examples: Bias in the training dataset, Lack of transparency in the model, Low diversity in the design team.

Impact Rating (**Severe**, Major, Moderate, Negligent).

Ecosystem Impact Rating:

Rating the impact of the violation on the ecosystem. Impact on the ecosystem includes potential harm to resources and ecosystems, harm to global supply chain or financial systems, and harm to natural resources and the planet (Tabassi, 2023).

Ecosystem Impact Rating scale examples:

Severe	Major	Moderate	Negligent
Resulting in long term, irreversible impact on the recourses, global supply chain, natural	Resulting in short term, irreversible impact on the recourses, global supply chain, natural	Resulting in short term, reversable impact on the resources, global supply chain, natural	Resulting in little to no impact on the resources, global supply chain, natural

resources, and ecosystems.	resources, and ecosystems.	resources, and ecosystems.	resources, and ecosystems.

These examples are meant to serve as possible descriptions of each category and are context dependent.

Organizational Impact Rating:

Rating the impact of the violation on the organization. Impact on the environment includes potential harm to business operations, security breaches, monetary loss, and harm to reputation (Tabassi, 2023).

Organizational Impact Rating scale examples:

Severe	Major	Moderate	Negligent
Resulting in permanent closure of the organization and irreversible reputational, legal, or financial setbacks.	Resulting in long-term down time/operations halt and long-term reputational, legal, or financial recovery.	Resulting in short term downtime/operations halt and reputational, legal, or financial recovery.	Little to no downtime and reputational, legal, or financial recovery.

These examples are meant to serve as possible descriptions of each category and are context-dependent.

People/Human Impact Rating:

Rating the impact of the violation on people. Harm to people includes harm to an individual's civil liberties and rights to physical and economic safety and security, harm to groups such as discrimination against a sub-group of a population, and societal violations such as harm to educational access, to civil rights, and democratic participation (Tabassi, 2023).

People/Human Impact Rating scale examples:

Severe	Major	Moderate	Negligent
Resulting in loss of human life, or permanent harm to human well-being (health, safety, security,	Resulting in long-term harm to human well-being (health, safety, security,	Resulting in short term reversible harm to human well-being (health, safety, security, privacy,	Resulting in no harm or short-term superficial harm that is reversable.

privacy, freedom, civil rights)	privacy, freedom, civil rights)	freedom, civil rights).	

These examples are meant to serve as possible descriptions of each category and are context dependent.

Value Violation Scenarios: Each value violation scenario represents a different type of scenario that could lead to the violation. This should represent a more detailed example of the possible scenarios that may lead to the value violation.
Examples: Lack of representation in the training data, No labels for protected classes in the training data, Insufficient protection protocols for personal identifying information.

Violation Path:
The violation path(s) represent the series of steps required for the value violation scenario to be executed. A value violation may have one or more possible paths.

Likelihood Rating:
The likelihood rating (Very High, High, Medium, Low, Very Low) represents the probability of the violation path happening. It is recommended to use the Likelihood Rating Score tool for this step.

Risk Score:
The risk score will be determined using the Risk Score Matrix, based on the highest Impact Rating and the highest Likelihood Rating as indicated in the examples.

Post-VARA

Once EVARA is concluded, depending on the results, it is recommended to consider the following:

- Transparency-based evaluation and protocols.
  Transparency aids in the openness and clarity of AI algorithms, processes, and outcomes.
  - The Open Ethics Data Passport is an example of a transparency-based evaluation tool for AI systems.
- Risk Mitigation Plans.
  Risk mitigation plans help with harm prevention, protection of stakeholder interests, compliance with regulations, and ethical alignment. Each plan is context-dependent.
- Reporting and Documentation.
  Robust documentation and reporting aid in ethical transparency, accountability, responsibility, and audit.
- System Reevaluation.
  It is important to reevaluate your system after a risk assessment until the desired outcome is reached. Risk assessment is often an iterative refinement process.

Nada Madkour nmadkour@emich.edu
March 9th 2024
GitHub

Use-Case Example(s)

Example 1: Self-Driving Car Pedestrian Detection

System Identification:
Define the target of this evaluation or the system under development.
Self Driving Car

Asset/Function Identification:
Define the asset or the function within the system being evaluated by E-VARA.
Pedestrian Detection Model

MLLC Phase Identification:
Define the asset MLLC phase. (Data Management, Training, Testing, Deployment, Use and Monitoring, Full cycle)
Data Management, Training

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

Chosen Value: Universal Usability: Freedom From Bias

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. Disparity in miss rates (between subgroups of pedestrians)	Negligible	Major	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	Steps
VS1. Inadequate representation in training data.	Path 1	Data is collected from an area with low diversity.
	Path 2	Data is collected from locations that are almost the same demographically, missing

		variations in pedestrian behavior and visual cues.
VS2. Inconsistent or inadequate labeling practices.	Path 1	Labels for protected classes are excluded from the training dataset.
	Path 2	Labeling practices are not uniform or standardized leading to significant variation.

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.

VV1VS1 - Possible value violation path	Likelihood
Path 1	
Path 2	
VV1VS2 - Possible value violation path	Likelihood
Path 1	
Path 2	

Risk Determination
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
VV1VS1	Very High	Severe	**6**
VV1VS2	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