Towards Multi-timescale Online Monitoring of AI Models: Principles and Preliminary Results

Fateh KAAKAI (Thales, IRT SystemX)
Paul-Marie RAFFI (IRT SystemX)

SafeAI 2023: The AAAI’s Workshop on AI Safety
February 13, 2023
Overview of the paper

- Why Monitoring
  - Rationale
  - Related Works

- How to Monitor
  - Methodological Aspects
  - Focus on Machine Learning

- Use Cases
  - Didactic examples
  - Welding conformity assessment (Renault)
Why Monitoring AI-based systems? … because many things can go wrong …

- The system shall **behave and shall continue to behave only as specified in the specified context**
- The system **does not behave or does not continue to behave as specified or in the specified context**

![Diagram](image)

**Certification Approval**

1. **Intent**
   - Insufficient Data Quality Attribute
   - Concept Drift

2. **Correctness**
   - Weak generalization capability
   - Lack of robustness
   - Intrinsic limitation of the AI technology

3. **Innocuity**
   - Lack of explainability
   - Unsafe unintended behavior

Source: Regulation (EU) 2017/373

Source: NASA/TM-2019-220292
How to design a good monitor?

Principle of the Multi-Timescale Monitoring

- **Data Collection**
  - Historical data
- **Offline Monitoring**
  - Monitor parameters
- **Online/ Runtime Monitoring**
  - Internal States
- **AI-based product**
- **Controller**
  - Operational data (e.g. sensors, logs)
  - Product output

![Diagram](image)

Trivial Regression Example

**AI-based Product Output**
- Expected behavior (ground truth)
- Output Robustness bounds
- Learning dataset
- ML Model

**Operating time**

- **Near-Past Monitoring (NPM)**
  - Issue detected
  - No issue detected
- **Present-Time Monitoring (PTM)**
  - Issue detected
  - No issue detected
- **Near-Future Monitoring (NFM)**
  - Issue detected
  - No issue detected
Monitoring the Operational Design Domain (ODD)

AI ML Model Performance Indicator (e.g., F-Score, MAE)

Specified Performance threshold

The ML Model underperforms in A with respect to specified MLCODD (lack of generalization of the model)

=> ODD PROBLEM REPORT for a limitation of use of the ML Model

MIN_BRIGHT as specified

The AI Model performs as expected with respect to the MLCODD (except locally in B and C)

=> ODD PROBLEM REPORT

B

Local decrease of ML Model Generalization

= Specified MLCODD (1)

Initial MLCODD as Specified

C

Local lack of ML Model Stability

= Final ML Constituent as Designed

Final ML Constituent as Designed

D

Opportunistic Model Robustness Zone

D

Out-of-Initial specified MLCODD (1)

= Out-of-Final designed MLCODD (2)

= Out-of-implemented MLMODD (3)

Out-of-Initial specified MLCODD (1)

Out-of-Final designed MLCODD (2)

Out-of-implemented MLMODD (3)

MAX_BRIGHT as specified

The AI Model performs as expected with respect to the specified MLCODD

=> ODD PROBLEM REPORT

=> ODD PROBLEM REPORT for a potential extension of the MLCODD

The ML Model overperforms in D with respect to the specified MLCODD

The ML Model is not robust, and it is not possible to claim any performance guarantee even in the opportunistic zone E

MIN_BRIGHT as specified

MAX_BRIGHT as specified

ODD parameter (e.g., level of image brightness)

System/Subsystem processes

ODD Problem Report

System/ subsystem requirements

ODD Problem Report

MLCODD process

System/Subsystem processes

MIN_BRIGHT as specified

MAX_BRIGHT as specified

The ML Model is not robust, and it is not possible to claim any performance guarantee even in the opportunistic zone E

System/Subsystem requirements

Monitoring the Operational Design Domain (ODD)
Illustration on Renault Welding Use Case
Key Engineering Principles to Design Online Monitor

- Monitoring Assets (Inputs, Outputs, State, Assumptions)
- Monitoring Time-Scale (NPS, PTM, NFM, Offline)
- Engineering approach (Top-Down, Bottom-Up, Mix)
- Monitoring Sophistication Proportionate to Risk (B-G-W boxes)

+ Independence of the Online Monitor
+ Performance of the Online Monitor
+ Innocuity of the Online Monitor
+ Hypervision of the Online Monitor
MultiTimescale Monitoring Poster

Visit us to discuss!