Province of Fryslân, Rijks ICT Gilde & the Z-Inspection® Initiative

Assessing the trustworthiness of an AI system in practice

Summary of results of the technical working group for the AI system "Monitoring grassification of heather fields"

Technology Readiness Level

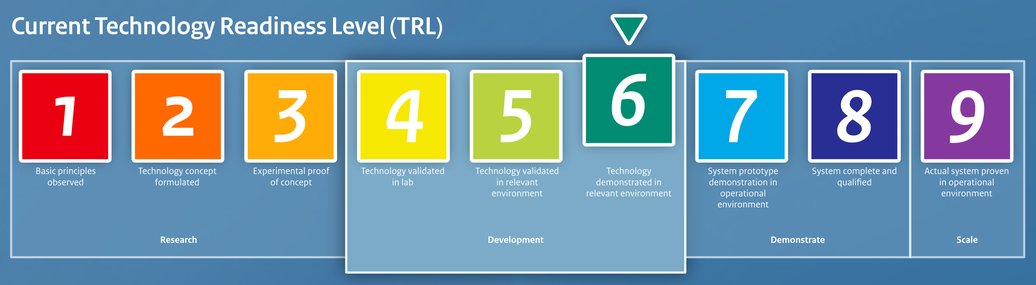
Assessing thrustworthy AI in practise

Findings

Route

1

Current Technology Readiness Level (TRL)



Status
of the algorithm

The technical maturity of the algorithm has been determined to be level 6, which means that the algorithm is under development and has not yet been tested in an operational environment. Level 6 is the starting point of the technical assessment. Each maturity level has its own focus points and associated research questions within the categories. This ensures that the algorithm is tested at the correct level appropriate to the current phase in the AI system life cycle.

2

Six major categories

The European Commission's HLEG-AI has developed a framework of ethics guidelines for trustworthy AI that rests on three pillars: ethically legitimate and robust. Based on this framework, six key categories have been identified for the assessment.

- Data management & processing
- Raw data & data labeling
- System & architecture
- Robustness
- Explainability
- Implementation & usage

3

Findings

Within these six categories, findings were identified that need to be worked on. The Province of Friesland can include these in a backlog to further develop the algorithm and advance it to the next technical readiness level. For each category, the top three findings are included in this summary.

Data management & processing

- Documentation for downloading and storing training and test data is missing.
- The model must be retrained periodically. How to do that and what people, knowledge and skills are needed to do that is still unclear.
- There is no pipeline yet for monitoring, retraining and redeploying the model. There is no machine learning operations (MLOps) process.

Raw data & data labeling

- There is relatively little training and testing data and this data may not be representative of the entire country or region.
- The quality of the training and testing data has not been quantified and measured, nor is it being monitored. Quality requirements and procedures need to be established and implemented.
- The ground truth needed to assess the accuracy of the model is not always fixed and may contain bias. It may lead to overestimation of model performance.

System & architecture

- No performance indicators have been established for monitoring and periodically evaluating the model in the operationalization and maintenance phase.
- A feedback process has not yet been established. Users cannot indicate which model outputs are incorrect, nor is this information used to improve the model.
- There is no common language to have a conversation about AI. Different words are used to mean the same thing, or words are not interpreted unambiguously.

Robustness

- It is not established how robust the model is. It is not clear what effect small changes in (the boundaries of) areas have on the outcome of the model.
- The model explanations have yet to be validated.
- It is difficult to make a good estimate of the robustness of the model because current satellite images are used in combination with possibly outdated ground truth labels.

Explainability

- End users need a lot of technical skills to try out the AI system and get a good idea of its performance.
- The meaning of various input variables and their impact on model output is not clear to end users.
- Visualizations of model outputs do not show where the model makes mistakes, while these are important starting points for further model improvement.

Implementation & usage

- It is not entirely clear how the model will eventually be used. How the model should be used and under what conditions it is valid has yet to be defined, for example in a model card.
- The training and testing dates overlap geographically. This can lead to bias in the model. The model should be tested on data from geographic areas that do not appear in the training set.
- Certain pixels in the satellite imagery have been identified as "not uniform" and are not included in the training and test data. This may lead to overestimation of model performance.

4

Reaching the next Technology Readiness Level (TRL)

Growth path
oh the algorithm

Step 3 describes the findings identified. To get the algorithm to the next technical readiness level, these findings must be addressed. Once the algorithm reaches level 7, a new assessment can be done, appropriate to that stage in the AI system life cycle.

Summaries assessment

The AI system was examined on three components: technical, ethical and ecological. The findings were captured in three different reports. A summary has been prepared for each report. In addition to substantive reports, lessons learned from applying the Z-inspection method were also identified. This is summarized in the lessons learned overview.

LESSONS LEARNED

TECHNOLOGY

ETHICS

ECOLOGY

Rijksorganisatie voor Ontwikkeling, Digitalisering en Innovatie
Ministerie van Binnenlandse Zaken en Koninkrijksrelaties

provinsje fryslân
provincie fryslân

Z-Inspection®