**Concepts of Design Assurance for NN (CoDANN)**

**March 31 2020**

**The review and scoring of assurance methods:**

Each elemental metric is allocated one point, and each method is either given that point or not (0 or 1), as follows:

* **I. Specificity to AI:** some assurance methods are generically tailored to many systems; others are deployable only to intelligent systems; one point was assigned to methods that focused (i.e. specifc) on the inner workings of AI systems.
* **II. The existence of a formal method:** this metric indicates whether the manuscript under review presented a formal (quantitative and qualitative) description of their method (1 point) or not (0 points).
* **III. Declared successful results:** in experimental work of a method under review, some authors declared success and presented success rates, if that is present, we gave that method a point.
* **IV. Datasets provided:** whether the method has a big dataset associated with it for testing (1) or not (0). This is an important factor for reproducibility and research evaluation purposes.
* **V. AI system size:** methods were applied to a small AI system, other were applied to bigger systems for instance, we gave a point to methods that could be applied to big real-world systems rather than ones with theoretical deployments.
* **VI. Declared success:** whether the authors declared success of their method in reaching an assured AI system (1) or not (0).
* **VII. Mentioned limitations:** whether there are obvious method limitations (0) or not (1).
* **VIII. Generalized to other AI deployments:** some methods are broad and are able to be generalized for multiple AI systems (1), others are “narrow” (0) and more specific to one application or one system.
* **IX. A real-world application:** if the method presented is applied to a real-world application, it is granted one point.
* **X. Contrasted with other methods:** if the method reviewed is compared, contrasted, or measured against other methods, or if it proves its superiority over other methods, then it is granted a point.

|  |  |
| --- | --- |
| **AIA Scoring Category** | **Why?** |
| **Specificity to AI (1)** |  |
| **The existence of a formal methods (1)** |  |
| **Declared successful results (1)** |  |
| **Datasets provided (1)** |  |
| **AI systems size (1)** |  |
| **Declared success (1)** |  |
| **Mentioned limitation (0)** |  |
| **Generalized to other AI deployments (1)** |  |
| **A real-world application (1)** |  |
| **Contrasted with other methods (1)** |  |

* Trade off between limitation vs generalization: If the methods is generalized then it certainly has limitation (exploration vs exploitation – similar to RL)

-----------------------------------------------------------------------------------------------------------------------------------------

**1/ Chapter 1 – Executive Summary:**

* Many concepts discussed in this report apply to machine learning algorithms in general, but an emphasis is put on the specific challenges of deep neural networks or deep learning for computer vision systems

**2/ Chapter 2 – Introduction:**

**a/ Background:**

* However, this increase of performance comes at the cost of more complexity in machine learning models, and this complexity might pose challenges in safety-critical domains, as it is often difficult to verify their design and to explain or interpret their behavior during operation

**b/ Learning Assurance Process Elements:**

Diagram

Description automatically generated

**3/ Chapter 3: Existing guidelines, standards and regulations, and their applicability to machine learning-based systems**

* All four building blocks are anticipated to have an importance in gaining confidence in the trustworthiness of an AI/ML application
  + **The AI trustworthiness analysis** should provide guidance to applicants on how to address each of the **seven key** guidelines in the **specific context of civil aviation**
  + The objective of Learning Assurance is to gain confidence at an appropriate level that an ML application supports the intended functionality, thus opening the **“AI black box”** as much as practically possible and required;
  + **Explainability** of AI is a human-centric concept that deals with the capability to explain how an AI application is coming to its results and outputs
  + **AI safety risk mitigation** is based on the anticipation that the “AI black box” may not always be opened to a sufficient extent and that supervision of the function of the AI application may be necessary.

Diagram

Description automatically generated

* Need building blocks for AI Trustworthiness of Data for policy maker.

EXAMPLE:

|  |  |  |  |
| --- | --- | --- | --- |
| EGTA key requirement | Constituting principles | Example of applicability to AI in safety-critical avionics | EASA building blocks |
|  |  |  |  |
|  |  |  |  |

Assurance (Data and Model verification techniques – does not have to be algorithms to dissect XAI or secure AI)

* **Requirements Management and Verification** are considered to be covered by traditional system development methods
* **Data Management:**
  + The data management process is the first step of the data life-cycle management:
  + considerations on the quality of the datasets
  + it should cover objectives on the independence between datasets and an evaluation of the bias and variance inherent to the data
* **Learning Process Management:**
  + It drives the **selection and validation** of key elements such as the **training algorithm, the activation function, the loss function, the initialization strategy, and the training hyperparameters**, which all have the potential to influence the result of the training in terms of performance.
  + Another consideration is on the **training environment**, including the host hardware and software frameworks, whose selection should be recorded and analyzed for potential risks.
* **Model Training:**
  + executing the training algorithm in the conditions defined in the previous step
  + . Once trained, the model performance, bias and variance are evaluated, using the validation dataset
* **Learning process verification:**
  + aims at evaluating the trained model performance on the **test dataset**. An evaluation of the bias and variance of the trained model should be performed.
* **Model Implemetation - deployment:**
  + The model implementation consists of transforming the training model into an executable model that can run on a target hardware.
* **Inference model verivication – evaluate performance on training and testing set:**
  + aims at verifying that the inference model behaves adequately compared to the trained model, by evaluating the model performance with the test dataset and explaining any differences in the evaluation metric compared to the one used in the training phase verification
* **Data verification:**