Verification and Validation of Systems in Which AI is a Key Element

Many systems are being considered in which artificial intelligence (AI) will be a key element. Failure of an AI element can lead to system failure (Dreossi et al 2017), hence the need for AI verification and validation (V&V). The element(s) containing AI capabilities is treated as a subsystem and V&V is conducted on that subsystem and its interfaces with other elements of the system under study, just as V&V would be conducted on other subsystems. That is, the high-level definitions of V&V do not change for systems containing one or more AI elements.

However, AI V&V challenges require approaches and solutions beyond those for conventional or traditional (those without AI elements) systems. This article provides an overview of how machine learning components/subsystems “fit” in the systems engineering framework, identifies characteristics of AI subsystems that create challenges in their V&V, illuminates those challenges, and provides some potential solutions while noting open or continuing areas of research in the V&V of AI subsystems.

Contents

Overview of V&V for AI-based Systems
Overview of V&V for AI-based Systems

Conventional systems are engineered via 3 overarching phases, namely, requirements, design and V&V. These phases are applied to each subsystem and to the system under study. As shown in Figure 1, this is the case even if the subsystem is based on AI techniques.

![Figure 1. Systems Engineering Phases for Systems Containing Machine Learning and Conventional Subsystems. (SEBoK Original, modeled after (Kuwajima et al. 2020))](image)

AI-based systems follow a different lifecycle than do traditional systems. As shown in the general machine learning life cycle illustrated in Figure 2, V&V activities occur throughout the life cycle. In addition to requirements allocated to the AI subsystem (as is the case for conventional subsystems), there also may be requirements for data that flow up to the system from the AI subsystem.
Characteristics of AI Leading to V&V Challenges

Though some aspects of V&V for conventional systems can be used without modification, there are important characteristics of AI subsystems that lead to challenges in their verification and validation. In a survey of engineers, Ishikawa and Yoshioka (2019) identify attributes of machine learning that make the engineering of same difficult. According to the engineers surveyed, the top attributes with a summary of the engineers’ comments are:

- **Lack of an oracle**: It is difficult or impossible to clearly define the correctness criteria for system outputs or the right outputs for each individual input.
- **Imperfection**: It is intrinsically impossible to for an AI system to be 100% accurate.
- **Uncertain behavior for untested data**: There is high uncertainty about how the system will behave in response to untested input data, as evidenced by radical changes in behavior given slight changes in input (e.g., adversarial examples).
- **High dependency of behavior on training data**: System behavior is highly dependent on the training data.

These attributes are characteristic of AI itself and can be generalized as follows:

- Erosion of determinism
- Unpredictability and unexplainability of individual outputs (Sculley et al., 2014)
- Unanticipated, emergent behavior, and unintended
consequences of algorithms

- Complex decision making of the algorithms
- Difficulty of maintaining consistency and weakness against slight changes in inputs (Goodfellow et al., 2015)

V&V Challenges of AI Systems

Requirements

Challenges with respect to AI requirements and AI requirements engineering are extensive and due in part to the practice by some to treat the AI element as a “black box” (Gunning 2016). Formal specification has been attempted and has shown to be difficult for those hard-to-formalize tasks and requires decisions on the use of quantitative or Boolean specifications and the use of data and formal requirements. The challenge here is to design effective methods to specify both desired and undesired properties of systems that use AI- or ML-based components (Seshia 2020).

A taxonomy of AI requirements engineering challenges, outlined by Belani and colleagues (2019), is shown in Table 1.

Table 1: Requirements engineering for AI (RE4AI) taxonomy, mapping challenges to AI-related entities and requirements engineering activities (after (Belani et al., 2019))

<table>
<thead>
<tr>
<th>RE Activities</th>
<th>AI Related Entities</th>
<th>System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elicitation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Availability of large datasets</td>
<td>- Lack of domain knowledge</td>
<td>- How to define problem/scopes</td>
</tr>
<tr>
<td>- Requirements analyst upgrade</td>
<td>- Undeclared consumers</td>
<td>- Regulation (e.g., ethics) not clear</td>
</tr>
<tr>
<td>Analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Imbalanced datasets, silos</td>
<td>- No trivial workflows</td>
<td>- No integration of end results</td>
</tr>
<tr>
<td>- Role: data scientist needed</td>
<td>- Automation tools needed</td>
<td>- Role: business analyst upgrade</td>
</tr>
</tbody>
</table>
Data is the life-blood of AI capabilities given that it is used to train and evaluate AI models and produce their capabilities. Data quality attributes of importance to AI include accuracy, currency and timeliness, correctness, consistency, in addition to usability, security and privacy, accessibility, accountability, scalability, lack of bias and others. As noted above, the correctness of unsupervised methods is embedded in the training data and the environment.

There is a question of coverage of the operational space by the training data. If the data does not adequately cover the operational space, the behavior of the AI component is questionable. However, there are no strong guarantees on when a data set is ‘large enough’. In addition, ‘large’ is not sufficient. The data must sufficiently cover the operational space.

Another challenge with data is that of adversarial inputs.
Szegedy et al. (2014) discovered that several ML models are vulnerable to adversarial examples. This has been shown many times on image classification software, however, adversarial attacks can be made against other AI tasks (e.g., natural language processing) and against techniques other than neural networks (typically used in image classification) such as reinforcement learning (e.g., reward hacking) models.

Model

Numerous V&V challenges arise in the model space, some of which are provided below.

- **Modeling the environment**: Unknown variables, determining the correct fidelity to model, modeling human behavior. The challenge problem is providing a systematic method of environment modeling that allows one to provide provable guarantees on the system’s behavior even when there is considerable uncertainty about the environment. (Seshia 2020)

- **Modeling learning systems**: Very high dimensional input space, very high dimensional parameter or state space, online adaptation/evolution, modeling context (Seshia 2020).

- **Design and verification of models and data**: data generation, quantitative verification, compositional reasoning, and compositional specification (Seshia 2020). The challenge is to develop techniques for compositional reasoning that do not rely on having complete compositional specifications (Seshia 2017).

- **Optimization strategy must balance between over- and under-specification**. One approach, instead of using distance (between predicted and actual results) measures, uses the cost of an erroneous result (e.g., an incorrect classification) as a criterion (Faria, 2018) (Varshney, 2017).

- **Online learning**: requires monitoring; need to ensure its exploration does not result in unsafe states.

- **Formal methods**: intractable state space explosion from complexity of the software and the system’s interaction with its environment, an issue with formal specifications.

- **Bias** in algorithms from underrepresented or incomplete training data OR reliance on flawed information that reflects historical inequities. A biased
algorithm may lead to decisions with collective disparate impact. Trade-off between fairness and accuracy in the mitigation of an algorithm’s bias.

- **Test coverage**: effective metrics for test coverage of AI components is an active area of research with several candidate metrics, but currently no clear best practice.

## Properties

Assurance of several AI system properties is necessary to enable trust in the system, e.g., the system’s trustworthiness. This is a separate though necessary aspect of system dependability for AI systems. Some important properties are listed below and though extensive, are not comprehensive.

- **Accountability**: refers to the need of an AI system to be answerable for its decisions, actions and performance to users and others with whom the AI system interacts
- **Controllability**: refers to the ability of a human or other external agent to intervene in the AI system’s functioning
- **Explainability**: refers to the property of an AI system to express important factors influencing the AI system results or to provide details/reasons behind its functioning so that humans can understand
- **Interpretability**: refers to the degree to which a human can understand the cause of a decision (Miller 2017)
- **Reliability**: refers to the property of consistent intended behavior and results
- **Resilience**: refers to the ability of a system to recover operations quickly following an incident
- **Robustness**: refers to the ability of a system to maintain its level of performance when errors occur during execution and to maintain that level of performance given erroneous inputs and parameters
- **Safety**: refers to the freedom from unacceptable risk
- **Transparency**: refers to the need to describe, inspect and reproduce the mechanisms through which AI systems make decisions, communicating this to relevant stakeholders.
V&V Approaches and Standards

V&V Approaches

Prior to the proliferation of deep learning, research on V&V of neural networks touched on adaptation of available standards, such as the then-current IEEE Std 1012 (Software Verification and Validation) processes (Pullum et al. 2007), areas need to be augmented to enable V&V (Taylor 2006), and examples of V&V for high-assurance systems with neural networks (Schumann et al., 2010). While these books provide techniques and lessons learned, many of which remain relevant, additional challenges due to deep learning remain unsolved.

One of the challenges is data validation. It is vital that the data upon which AI depends undergo V&V. Data quality attributes that are important for AI systems include accuracy, currency and timeliness, correctness, consistency, usability, security and privacy, accessibility, accountability, scalability, lack of bias, and coverage of the state space. Data validation steps can include file validation, import validation, domain validation, transformation validation, aggregation rule and business validation (Gao et al. 2011).

There are several approaches to V&V of AI components, including formal methods (e.g., formal proofs, model checking, probabilistic verification), software testing, simulation-based testing and experiments. Some specific approaches are:

- Metamorphic testing to test ML algorithms, addressing the oracle problem (Xie et al., 2011)
- A ML test score consisting of tests for features and data, model development and ML infrastructure, and monitoring tests for ML (Breck et al., 2016)
- Checking for inconsistency with desired behavior and systematically searching for worst-case outcomes when testing consistency with specifications.
- Corroborative verification (Webster et al., 2020), in which several verification methods, working at different levels of abstraction and applied to the same AI component, may prove useful to verification of AI components of systems.
- Testing against strong adversarial attacks (Useato, 2018); researchers have found that models may show
robustness to weak adversarial attacks and show little to no accuracy to strong attacks (Athalye et al., 2018, Uesato et al., 2018, Carlini and Wagner, 2017).

- Use of formal verification to prove that models are consistent with specifications, e.g., (Huang et al., 2017).

- Assurance cases combining the results of V&V and other activities as evidence to support claims on the assurance of systems with AI components (Kelly and Weaver, 2004; Picardi et al. 2020).

## Standards

Standards development organizations (SDO) are earnestly working to develop standards in AI, including the safety and trustworthiness of AI systems. Below are just a few of the SDOs and their AI standardization efforts.

ISO is the first international SDO to set up an expert group to carry out standardization activities for AI. Subcommittee (SC) 42 is part of the joint technical committee ISO/IEC JTC 1. SC 42 has a working group on foundational standards to provide a framework and a common vocabulary, and several other working groups on computational approaches to and characteristics of AI systems, trustworthiness, use cases, applications, and big data. ([https://www.iso.org/committee/6794475.html](https://www.iso.org/committee/6794475.html))


Underwriters Laboratory has been involved in technology safety for 125 years and has released ANSI/UL 4600 “Standard for Safety for the Evaluation of Autonomous Products”. ([https://ul.org/UL4600](https://ul.org/UL4600))

The SAE G-34, Artificial Intelligence in Aviation, Committee is responsible for creating and maintaining SAE Technical Reports, including standards, on the implementation and certification aspects related to AI technologies inclusive of any on or off-board system for the safe operation of aerospace systems and aerospace vehicles. ([https://www.sae.org/works/committeeHome.do?comtID](https://www.sae.org/works/committeeHome.do?comtID))
References

Works Cited


Gunning, D. Explainable Artificial Intelligence (XAI). In IJCAI 2016 Workshop on Deep Learning for Artificial Intelligence (DLAI), July 2016.


Conference on Artificial Intelligence (AAAI-20) (2020)


Szegedy, Christian, Zaremba, Wojciech, Sutskever, Ilya,


Primary References


**Additional References**


Raj, S., L.L. Pullum, A. Ramanathan, and S.K. Jha,


