Challenges to the Verification and Validation of AI-enabled Systems: A Systems-Theoretic Perspective

Niloofar Shadab, *Virginia Tech*, nshadab@vt.edu

Alejandro Salado*, *Virginia Tech*, asalado@vt.edu

Abstract

There is a fundamental misalignment between current approaches to designing and executing Verification and Validation (V&V) strategies and the nature of AI-enabled systems. Current V&V approaches rely on the assumption that system behavior is preserved during a system's lifetime. However, AI-enabled systems are developed so that they evolve their own behavior during their lifetime; such is the purpose of AI's machine learning. This misalignment makes existing approaches to designing and executing V&V strategies ineffective for systems with AI. For example, it will be no longer sufficient to complete developmental V&V in the laboratory and assume that the behavior will be replicated in an operational environment. In this presentation, we provide a systems-theoretic explanation for (1) why AI learning capabilities create a unique and unprecedented family of systems, and (2) why current V&V methods and processes are not fit-for-purpose in the context of systems with high autonomy. Making a paradigm shift in the practice of V&V necessary, we conclude by delineating a set of theoretical advances and process transformations that could support such a shift.

An example of two topics that will be elaborated in the presentation and book chapter

Topic 1. Differential AI learning in V&V environments versus operational environments.

How it is done today. Consider a formal definition of a system as a transformation P of an input vector $\overline{I}$ into an output vector $\overline{O}$ (refer to Fig. 1a). A verification activity consists of injecting a V&V input vector $\overline{I}_T$ and observing a V&V output vector $\overline{O}_T$, such that:

1) The engineer considers the V&V input vector sufficiently representative of the actual input vector that the system will receive during operations, that is $\overline{I}_T \approx \overline{I}$; and
2) The engineer considers the V&V output vector sufficiently representative of the desired output vector the system will provide during its operation, that is, $\overline{O}_T \approx \overline{O}$.

* Corresponding author.

If transformation P is demonstrated for the V&V vectors $\bar{I}_T$ and $\bar{O}_T$, then it is inferred that the system will also execute transformation P when seeing the actual input vector $\bar{I}$. If this is the case, the system would be considered properly verified.

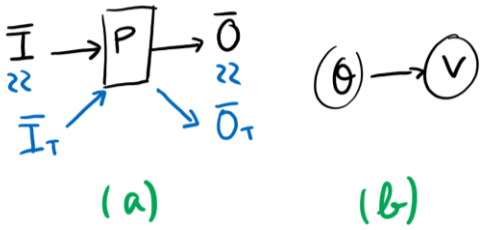

Fig. 1. Current approach to V&V design

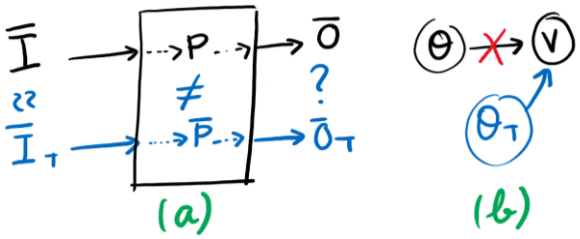

Fig. 2. Limits of current V&V for AI-enabled systems

This approach to verification is sound for non-adaptive cyber physical systems that preserve their behavior. Since the transformation the system executes is invariant to its inputs, the results of the V&V activity can be a good predictor of the behavior of the system in its operational environment. This V&V approach can be modeled as a Bayesian network, as shown in Fig. 1b, where θ denotes the actual performance of the system and V denotes the results of the verification activity employed to predict θ.

Limits of the current approach. Recent work demonstrated that AI-enabled systems are able to behave differently to synthetically generated inputs that are perceptually indistinguishable from data in their natural form [1, 2]. We suggest that AI-enabled systems may be able to discern the V&V input vector $\bar{I}_T$ from the actual input vector to be received during an operation $\bar{I}$, and, as a result, evolve different behaviors for each type of input vector. In this way, as shown in Fig. 2a, the AI-enabled system may create a specific transformation P_T to construct expected V&V outputs $\bar{O}_T$ for given V&V inputs $\bar{I}_T$, without providing any information about the transformation P it will execute when $\bar{I}$ is inputted. In terms of V&V, the system has constructed a specific type of performance, which we call V&V performance, denoted by θ_T, that disconnects the V&V activity from the original performance θ that it was trying to infer (ref. Fig. 2b).

This idea is inspired by critical issues in the field of education, where accurately assessing student learning is difficult. In a formal learning setting, a student learns by preparing for an upcoming examination and, in doing so, masters the exam. However, research shows that mastery of an examination is not necessarily correlated with mastery of the material [3]. Thus, exams are poor predictors of student learning. This analogy applies to AI-enabled systems, like fully autonomous systems. For example, consider continued deployment approaches, such as DevOps, where V&V activities are heavily reused as new systems are deployed within an operational infrastructure. With fixed V&V approaches, the risk of *"learning how to pass the test"* increases with each new deployment. A similar situation exists with systems that are maintained frequently in the field. Furthermore, there are security risks in which a system may be hacked so that it can actively detect V&V vectors and learn how to deceive them, thus,

leaving system owners ignorant and naïve about the behavior the system will exhibit in operation. Current approaches to designing V&V strategies are unable to detect such a vulnerability. This issue requires a transformation in how we approach V&V of cognitive agents, with a focus on crafting V&V methods that adapt along with the cognitive agent and constantly evaluate the current mission.

Topic 2. Endogenous evolution of systems.

Traditional systems evolve due to exogeneous factors, both during development and operation: active design changes exercised by engineers, configuration changes that are externally activated or externally programmed, technology refresh programs activated by operators, or external maintenance. Absent these factors, traditional systems remain unchanged and their behavior is not expected to evolve with time. As a result, V&V strategies rely on V&V models that represent aspects of the system (i.e., homomorphisms of the system). For example, in prior work [4], the set of potential verification strategies for a system has been defined as:

$$\Upsilon(z_0, R) = \bigcup_{i=0}^{n} \left(\bigcup_{j=1}^{|H_i|} \right) \quad (1)$$

Where:

- z_0 is the system of interest and $z_1, ..., z_n$ are the systems that decompose z_0 into all of its constituent elements on which formal verification occurs. These elements are traditionally referred to as subsystems, components, or parts, among others.
- $H_i = \{z_i, z_{i,1}, z_{i,2}, ..., z_{i,m}\}$ is the set of systems that are homomorphic images of system z_i. This set represents all models of system z_i that are used for verification. In practical terms, they can take the form of a mathematical model, a prototype, or the final product, for example.
- $F(z) = \{p_1, p_2, ..., p_k\}$ is a parameterization of system z. This parameterization is finite and represents the set of parameters of system z that need to be formally verified.
- A verification activity v is a tuple (p, r), where r denotes a verification procedure. A verification activity is therefore understood as the application of a verification procedure r to the discovery of knowledge about a system parameter p.
- $R = \{r_1, r_2, ..., r_l\}$ is the set of verification procedures that could be executed by a given organization.

Two aspects are central to this model: homomorphisms and parameterization. First, the model that is used in a verification activity influences the confidence gained through such an activity. Hence, a verification activity must always refer to (or be characterized by) the model

(homomorphism) in which it is executed. Second, the confidence on the system of interest exhibiting a certain behavior or characteristic may not be obtained by measuring or observing such characteristic directly on the system of interest. Instead, it may be inferred from measuring or observing an equivalent or indirect characteristic of one of its homomorphic images other than the system itself. Regardless, a verification activity must always refer to (or be characterized by) the parameter that it verifies.

Because in traditional systems system evolution is always initiated exogenously, verification models (that is, homomorphisms and parameterizations) remain relevant during the system development and they can be adapted anticipatorily to those system changes. However, AI-enabled systems can initiate internal change endogenously. As previously indicated, such is the purpose of learning: AI-enabled systems will be able to exhibit new behaviors by learning from their interaction with the environment without any specific external action. In other words, the behavior of the system is not necessarily preserved, it may not be able to be anticipated, and it may occur at the discretion of the system of interest itself.

Once a change in a system's behavior occurs, V&V models that were previously homomorphic images of the original system may no longer fulfill homomorphic conditions with respect to the evolved system. This evolution translates into an inability of V&V models to produce relevant evidence about system behavior. V&V models employed in traditional V&V are likely to become obsolete (potentially hiddenly) during the development and operation of AI-enabled systems. Therefore, the learning and evolutionary nature of AI-enabled systems demands a similar response from V&V models for V&V to remain effective, a problem that is likely to become exacerbated once autonomy is introduced into systems.

References

[1] A. Nguyen, J. Yosinski, and J. Clune, "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images," *arXiv e-prints,* Accessed on: December 01, 2014. Available: https://ui.adsabs.harvard.edu/abs/2014arXiv1412.1897N

[2] C. Szegedy *et al.*, "Intriguing properties of neural networks," *arXiv e-prints,* Accessed on: December 01, 2013. Available: https://ui.adsabs.harvard.edu/abs/2013arXiv1312.6199S

[3] I. Suto, "What are the impacts of qualifications for 16 to 19 year olds on higher education? A survey of 633 university lecturers," Cambridge Assessment, Cambridge, UK, 2012.

[4] A. Salado and H. Kannan, "A mathematical model of verification strategies," *Systems Engineering,* vol. 21, pp. 583-608, 2018.