The Path to Consensus on Artificial Intelligence Assurance

Laura Freeman
Hume Center for National Security and Technology
Virginia Tech
Arlington, VA, USA
laura.freeman@vt.edu

Feras A. Batarseh
Bradley Department of Electrical and Computer Engineering
Virginia Tech
Arlington, VA, USA
batarseh@vt.edu

D. Richard Kuhn
National Institute of Standards and Technology (NIST)
Gaithersburg, MD, USA
kuhn@nist.gov

M S Raunak
National Institute of Standards and Technology (NIST)
Gaithersburg, MD, USA
raunak@nist.gov

Raghu N Kacker
National Institute of Standards and Technology (NIST)
Gaithersburg, MD, USA
raghu.kacker@nist.gov

Widescale adoption of intelligent algorithms requires that Artificial Intelligence (AI) engineers provide assurances that an algorithm will perform as intended. Providing such guarantees involves quantifying capabilities and the associated risks across multiple dimensions including: data quality, algorithm performance, statistical considerations, trustworthiness, security, as well as explainability. In this article we discuss the formalization of important aspects of AI assurance, including its key components.

I. The State of AI Assurance

In recent years, there has been a renewed focus on the field of AI Assurance. Researchers, policymakers, and business leaders all use the phrase AI assurance, but there is little consensus on what this term precisely means. Batarseh et al. define AI assurance as: "a process that is applied at all stages of the AI engineering lifecycle ensuring that any intelligent system is producing outcomes that are valid, verified, data-driven, trustworthy and explainable to a layman, ethical in the context of its deployment, unbiased in its learning, and fair to its users" [1].

As nations race to develop AI, lack of attention to assurance is giving rise to some serious concerns. Given the availability of the technology and the potentially massive economic benefits, progress in AI will inevitably bring about a race to AI assurance. It is thus critical to define a consensus-based path forward.

The European Commission (EC) of the European Union (EU) recently proposed rules and actions for excellence and trust in AI systems across the continent [2]. The new AI legal framework aims to ensure that Europeans can trust AI systems across all domains, and address the specific risks posed by AI systems. Achieving trust and confidence in AI systems will require international consensus. This article aims to raise issues and promote discussion in this critical area.

A. The Need for an AI Assurance Discipline

The assurance of AI systems has been an Achilles heel up until now. A parallel was witnessed in general consumer software, where systematic and measurable testing throughout the lifecycle was often an afterthought. Learning from that experience, we should not treat assurance as a separate component. Rather, assurance should be a part of the incremental learning process of any intelligent agent, algorithm, or environment. In AI development however, a significant gap is observed, one that exists between AI systems' abilities to generalize from their learning domains (i.e., data) to their ability to create a credible view of the world (i.e., their operating environment). If a model learns from features unique to the training domain, but not observed in the broader world, it creates patterns that are not an assured reflection of their context. This leads to the AI system losing its ability to make accurate predictions, recommendations, or classifications, especially in a new environment. Accordingly, data assurance (of the input data used in training and testing) and algorithmic assurance are both equally critical to the adoption of AI systems.

B. Key Aspects of AI Assurance

AI systems differ from conventional software in many ways, and these differences require different approaches for assurance:

1) **Verification & Validation of AI**: Software testing activities involve two main aspects: verification and validation, often referred to as V&V. Verification checks

if the system is being build right, i.e., without errors or defects, while validation means providing the desired system to the user; it is building the right system. Like conventional software, assuring AI models requires performing V&V activities, but it expands beyond those limits towards the evaluation of learning, inputs to the algorithms, data quality, and other environmental aspects that play a major role (such as fairness, context, and ethics). One of AI assurance's aspects that is fairly novel (and is only relevant to AI) is Explainable AI (XAI) [3]. Understanding and interpreting AI algorithms (albeit a difficult task) is critical to their adoption.

2) **AI System Development & Deployment**: Intelligent systems are being deployed everywhere, but every domain has unique considerations in defining their scope. A system deployed at a hospital for instance, requires a different type of assurance than an intelligent system flying a fighter jet. Moreover, the increasing use of learning algorithms in cyberphysical systems (such as smart grids, autonomous transportation, and smart farming) has pushed the need for ongoing security and safety checks. The deployed AI system hence requires a structured sequence of tests across the development and deployment lifecycles coupled with statistical analyses of the data and the models' outcomes. These tests include the selection of test data sets to test the AI algorithm itself and tests of the deployed AI-enabled system in the deployed environment [4].
3) **Transparency**: Users of AI systems "have a right" to have outputs and decisions affecting them explained in an understandable way, and preferably using domain-specific terms and formats. Besides increasing their trust in the system, this allows domain experts as well as regular users to inspect the system's processes and manage its goals.
4) **Bias**: Bias can be statistical (i.e., detectable through overfitting and underfitting measures) or due to issues such as skew or incomplete data in "the environment." Bias can be be investigated and mitigated through data collection best practices, the analysis of contextual awareness, statistical measures (e.g., Q-value), analysis of variance (ANOVA) method including lack of fit analysis, outlier detection methods such as isolation forests, causal inference, and other similarity matrices. Other data engineering practices affect bias, for instance, data collection ambiguities, data imbalance and availability, and the lack of domain information.
5) **Context**: Often refered to as situational awareness, context across domains can change drastically. An AI agent for online misinformation detection aims to ensure fairness and trustworthiness values, while an AI agent regulating a nuclear reactor prioritizes safety and security. Contextual modeling is a difficult problem; to "capture" context, knowledge elicitation is performed (a derivative of requirements engineering), as well as the scoping and collection of dark data (data that could be available but are not clearly correlated with the domain [5]), and leveraging data fusion to unify datasets from multiple sources. The difference in goals across domains is addressed through capturing and modeling context and context-based reasoning

From a theoretical and a practical perspective, the key aspects mentioned in this section are essential to the accomplishment of AI assurance goals, which are introduced next.

II. AI Assurance Goals

Users often do not trust, adopt, or use algorithmic systems if they do not understand how they work. Studies suggest that scientists (and the public) would be much more willing to accept algorithmic decision-support if explainability, trustworthiness, and other assurance measures are provided [6]. Accordingly, we make the case that assurance is required to enable the adoption of AI - and the overall avoidance of an AI bubble burst. However, similar to any other engineering discipline, analyzing trade-offs between engineering goals is both an art and a science. Due to the recent emergence of big data, AI assurance manifested in different forms, not limited to only the verification and validation of the algorithm, but also to the quality of data, as well as philosophical challenges like dark data, causality, and bias. Accordingly, issues such as breach of ethics by AI systems lead to the need for "loading" goals into the system to allow it to learn things "correctly," creating a challenge in formulating qualitative measures (such as ethics) into an AI system. Based on the survey by Batarseh, Freeman, and Huang [1], the following six goals are extracted from literature as the main goals relevant to AI assurance: (1) eXplainable AI, (2) Safe AI, (3) Secure AI, (4) Trustworthy AI, (5) Ethical AI, and (6) Fair AI. All of these goals are challenges that require a dedicated set of solutions and methods by domain and by AI approach, although model- and domain-agnostic assurance approaches are also viable potentials.

AI is often assessed by its ability to consistently deliver accurate predictions of behavior in a system. A critical, often overlooked, aspect of developing AI algorithms is that performance is a function of the task the algorithm is assigned, the domain over which the algorithm is intended to operate, and the changes to these elements in time. These parameters and their constituent parts form the basis over which assuring AI becomes a challenge. Algorithms need to be characterized by understanding the factors that contribute to stable performance across an operational environment.

To accurately and consistently predict outputs or behaviors, AI enabled systems require data for training, testing, as well as validating predictions. The iterative process of improving accuracy and precision in developed models involves trade-offs in performance, data quality, and other environmental factors. AI's predictive power can be impacted through changes in the training or test data, the model, and the environment. In this section, we discuss sources of change captured within the operational envelope of an AI system's execution, which is often attributed to their inconsistencies. Model and data changes have been discussed in the literature around concept

drift and are examples of how these inconsistencies could be measured.

Even for conventional software, assurance is difficult to quantify, and often the best measures that can be provided involve levels of structural coverage. For example, life-critical aviation software must pass requirements-based tests that provide 100% Modified Condition Decision Coverage (MCDC). But structural coverage criteria such as MCDC or branch coverage do not apply to neural networks or other AI approaches, which are often 'programmed' through the inputs. That is, the accuracy of the AI model is almost completely dependent on the data set used in training. The key question, then, is whether the training inputs represent the real world in sufficient depth, and we want to quantify this assurance.

One approach for quantifying input model adequacy for deep learning algorithms is to evaluate 'neuron coverage', i.e., the fraction of neurons that are exercised during training and testing. However, empirical data on the effectiveness of this metric suggest only weak correlations with test effectiveness [7]. More recently, we have investigated combinatorial coverage measurement and differencing. This method computes the fraction of t-way combinations of parameter values included in inputs. The intuition is that *combinations* of inputs are essential to accuracy in AI algorithms, so measuring thoroughness of combination coverage provides valuable information on input or training data adequacy. Initial evidence has shown this method to be effective for evaluating the quality of input models and transfer learning data models [8]. These concepts are applicable to other AI approaches such as reinforcement learning and genetic algorithms.

III. Achieving AI Assurance

To accomplish the six assurance goals, along with other recommendations, a foundational consensus for defining and measuring the dependability of AI Systems is needed.

A. A Community Consensus for AI

AI systems developed and deployed by different researchers, organizations, and government agencies are likely to have a wide range of maturity in terms of providing assurance. A well articulated process and a clearly defined set of metrics to categorize and evaluate these systems could go a long way in establishing a common understanding of these systems' dependability. The advantage of such a consensus evaluation model is that it encourages all stakeholders to agree on a set of metrics and processes to measure the quality of the AI systems being produced and deployed. It also shows the path to achieve a gradually higher level of assurance following a consensus set of criteria. We believe that community agreement on a similar set of metrics and process will not only streamline AI systems' development efforts, it will also foster sharing of implementation experiences and best practices.

AI strategy has become an international priority. Countries like the US, Japan, the United Kingdom, Russia, Germany, and China are just a few of the dozens of countries that have issued national strategies around AI, see for example [9] and [10]. In the United States, as organizations are evaluating the use of AI, they have developed different implementation strategies, guiding principles, and ethics statements. For example, in 2019, the U.S. AI Initiative summarized American AI values in four different aspects: Understandable and Trustworthy AI, Robust and Safe AI, Workforce Impact, and International Leadership. All these aspects are well-grounded in assurance. Similarly, multiple research projects are underway at NIST directly and indirectly shaping the different aspects of AI assurance. Many researchers at NIST are actively working on developing metrics, measurements, and tools for building and analyzing AI systems that are accurate, reliable, safe, secure, robust, explainable, privacy preserving, and free from bias. A community consensus for evaluating AI systems would be a vehicle to accomplish the mentioned goals.

B. Discussions and Future Goals

As we contemplate the success of learning algorithms, it is clear that extrapolating general intelligence or wide AI adoption will require consensus on the rules governing it, as well as its overall assurance. For that, we present the following conceptual considerations:

1) Assurance should not be an afterthought, rather, it should be embedded into the lifecycle of development and learning in all AI systems. Recent developments such as surrogate models constitute a positive development towards achieving incremental assurance.
2) Current AI models are almost exclusively statistical, they don't have the ability to grasp or represent context, we deem this aspect critical to the future of AI and its assurance.
3) Consider counterfactual scenarios for AI algorithms; at the end of the day, if AI algorithms cannot explain cause and effect, they may be rendered obsolete by the next big technology.
4) It has been suggested that a system validating a learning algorithm will be as complex as the learning system. Research and development of AI assurance approaches thus needs to receive attention comparable to AI applications research.

The future of AI certainly is promising, yet likely to be different than its recent past. Notions such as contextual adaptations and XAI will become more evident and dominant. Nonetheless, one aspect that all phases of AI require is assurance. For AI to reach its scientific and practical goals, and for humans to reap its benefits, AI researchers, practitioners, and investors need to be on a mission to display the virtuous goodness of a fair, safe, secure, explainable, trustworthy, and ethical AI.

Disclaimer: The views in this article are not official statements from the NIST AI Program, where extensive work in the area of Trustworthy AI is being conducted by multiple research groups.

REFERENCES

[1] F. A. Batarseh, L. Freeman, and C.-H. Huang, "A survey on artificial intelligence assurance," *Journal of Big Data*, vol. 8, no. 1, pp. 1–30, 2021.

[2] T. E. Union, "Proposal for a regulation of the european parliament and of the council laying down harmonised rules on ai (ai act)," https://ec.europa.eu/commission/presscorner.

[3] M. S. Raunak and R. Kuhn, "Explainable artificial intelligence and machine learning," *IEEE Computer*, vol. 54, no. 10, pp. 25–27, 2021.

[4] E. Lanus, I. Hernandez, A. Dachowicz, L. Freeman, M. Grande, A. Lang, J. H. Panchal, A. Patrick, and S. Welch, "Test and evaluation framework for multi-agent systems of autonomous intelligent agents," *arXiv preprint arXiv:2101.10430*, 2021.

[5] D. Hollister, "Data democracy for psychology: how do people use contextual data to solve problems and why is that important for ai systems?" in *Data Democracy*. Elsevier, 2020, pp. 163–177.

[6] E. Lanus, I. Hernandez, A. Dachowicz, L. Freeman, M. Grande, A. Lang, J. H. Panchal, A. Patrick, and S. Welch, "Test and evaluation framework for multi-agent systems of autonomous intelligent agents," *arXiv preprint arXiv:2101.10430*, 2021.

[7] J. R. Toohey, M. S. Raunak, and D. Binkley, "From neuron coverage to steering angle: Testing autonomous vehicles effectively," *Computer*, vol. 54, no. 8, pp. 77–85, 2021.

[8] E. Lanus, L. J. Freeman, D. R. Kuhn, and R. N. Kacker, "Combinatorial testing metrics for machine learning," in *2021 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*. IEEE, 2021, pp. 81–84.

[9] G. Allen, "Understanding china's ai strategy, center for a new american security"," available at: https://www.cnas.org/publications/reports/understanding-chinas-ai-strategy.

[10] White House, "Select committee on ai," available at: https://trumpwhitehouse.archives.gov/wp-content/uploads/2021/01/Charter-Select-Committee-on-AI-Jan-2021-posted.pdf.