The Path to a Consensus on Artificial Intelligence Assurance

Laura Freeman and Feras A. Batarseh, Virginia Tech

D. Richard Kuhn, M S Raunak, and Raghu N Kacker, National Institute of Standards and Technology

Wide-scale adoption of intelligent algorithms requires artificial intelligence (AI) engineers to provide assurances that an algorithm will perform as intended. In this article, we discuss the formalization of important aspects of AI assurance, including its key components.

Digital Object Identifier 10.1109/MC.2021.3129027
Date of current version: 11 March 2022

To ensure the wide-scale adoption of intelligent algorithms, artificial intelligence (AI) engineers must offer assurances that an algorithm will function as intended. Providing such guarantees involves quantifying capabilities and the associated risks across multiple dimensions, including data quality, algorithm performance, statistical considerations, trustworthiness, and security as well as explainability.

THE STATE OF AI ASSURANCE

In recent years, there has been a renewed focus on the field of AI assurance. Researchers, policy makers, 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[1]

> *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*

0018-9162/22©2022IEEE

EDITOR **HSIAO-YING LIN**
IEEE Member; hsiaoying.lin@gmail.com

driven, trustworthy, and explainable to a layman; ethical in the context of its deployment; unbiased in its learning; and fair to its users.

As nations race to develop AI, a lack of attention to assurance is giving rise to some serious concerns. Given the availability of technology and 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 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 an international consensus. This article aims to raise issues and promote discussion in this critical area.

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 (namely, data), to their ability to create a credible view of the world (that is, 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.

Key aspects of AI assurance

AI systems differ from conventional software in many ways. In the following sections, we discuss the different approaches required for assurance.

Verification and validation of AI. Software-testing activities involve two main aspects: verification and validation (V&V). Verification checks whether the system is being built right, namely, without errors or defects, while validation means providing the desired system to the user, that is, building the right system. Like conventional software, assuring AI models requires performing V&V activities, but it expands beyond those limits toward 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.

AI system development and 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]

> Achieving trust and confidence in AI systems will require an international consensus.

Transparency. The users of AI systems have a right to have outputs and decisions affecting them explained in an understandable way, 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.

Bias. Bias can be statistical (namely, detectable through overfitting and underfitting measures) or due to issues such as skew or incomplete data in the environment. Bias can be investigated and mitigated through data collection best practices; the analysis of contextual awareness; statistical measures (for example, Q-values); an analysis of variance methods, including lack-of-fit analysis; and outlier detection methods such as isolation forests,

causal inference, and other similarity matrices. Other data engineering practices affect bias, for example, data collection ambiguities, data imbalance and availability, and the lack of domain information.

Context. Often referred 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 data sets 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.

> AI's predictive power can be impacted through changes in the training or test data, the model, and environment.

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 tradeoffs between engineering goals is both an art and a science. Due to the recent emergence of big data, AI assurance has manifested itself in different forms, limited not only to the V&V 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 a 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 conducted by Batarseh et al.,[1] the following six goals are extracted from the literature as the main goals relevant to AI assurance: 1) XAI, 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 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 to which the algorithm is assigned, the domain over which the algorithm is intended to operate, and the changes to these elements over 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 and testing as well as validating predictions. The iterative process of improving accuracy and precision in developed models involves tradeoffs 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 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 its 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 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, that is, the fraction of neurons 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.

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. In the following sections, we discuss these consensus objectives in greater detail.

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 used 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 processes will not only streamline AI systems' development efforts, it will also foster the sharing of implementation experiences and best practices.

One motivation for consensus approaches is the need for context- and domain-specific assessments. The EC recently drafted the first-ever legal framework on AI: the Artificial Intelligence Act. This AI act (draft) has the potential to foster a community consensus because it advocates for harmonized rules and a tailorable, risk-based framework. This tailoring depends on the application and its associated risks.[2] Additionally, "Tools for Trustworthy AI" notes the need for comparing tools and practices for achieving trustworthy AI in the context of its application.[9] The consensus methods provide a tailorable methodology that can accommodate a context-specific application of AI.

AI strategy has become an international priority, which adds another layer of context. Countries including the United States, Japan, the United Kingdom, Russia, Germany, and China are just a few of the dozens of countries that have issued national strategies around AI.[10,11] Each country is identifying the specific priorities that need to be factored into the AI-assurance process. As organizations are evaluating the use of AI, they have developed different implementation strategies, guiding principles, and ethics statements. 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 of these aspects are well grounded in assurance. Similarly, multiple research projects are underway at the National Institute of Standards and Technology (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 aforementioned goals.

Discussions and future goals

As we contemplate the success of learning algorithms, it is clear that extrapolating general intelligence or wide-scale AI adoption will require a 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 toward achieving incremental assurance.
2. Current AI models are almost exclusively statistical, that is, 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. The R&D of AI-assurance approaches thus needs to receive attention comparable to AI applications research.

The future of AI is certainly promising yet likely to be different from 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

DISCLAIMER

The views expressed in this article are not official statements from NIST's AI program, where extensive work in the area of trustworthy AI is being conducted by multiple research groups.

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.

REFERENCES

1. F. A. Batarseh, L. Freeman, and C.-H. Huang, "A survey on artificial intelligence assurance," *J. Big Data*, vol. 8, no. 1, pp. 1–30, 2021, doi: 10.1186/s40537-021-00445-7.
2. "Proposal for a regulation of the European parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts," European Commission, Brussels. Accessed: 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206
3. M. S. Raunak and R. Kuhn, "Explainable artificial intelligence and machine learning," *Computer*, vol. 54, no. 10, pp. 25–27, 2021, doi: 10.1109/MC.2021.3099041.
4. E. Lanus *et al.*, "Test and evaluation framework for multi-agent systems of autonomous intelligent agents," 2021, *arXiv:2101.10430*.
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*, New York, NY, USA: Elsevier, 2020, pp. 163–177.
6. B. M. Keneni *et al.*, "Evolving rule-based explainable artificial intelligence for unmanned aerial vehicles," *IEEE Access*, vol. 7, pp. 17,001–17,016, Jan. 2019, doi: 10.1109/ACCESS.2019.2893141.
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, doi: 10.1109/MC.2021.3079921.
8. E. Lanus, L. J. Freeman, D. R. Kuhn, and R. N. Kacker, "Combinatorial testing metrics for machine learning," in *Proc. 2021 IEEE Int. Conf. Softw. Testing, Verification Validation Workshops (ICSTW)*, pp. 81–84, doi: 10.1109/ICSTW52544.2021.00025.
9. OECD, 2021, "Tools for trustworthy AI: A framework to compare implementation tools for trustworthy AI systems," OECD, Digital Economy Papers, No. 312, OECD Publishing, Paris. [Online]. Available: https://doi.org/10.1787/008232ec-en
10. G. Allen, "Understanding China's AI strategy," Center for a New American Security, Washington, DC, USA, 2019. [Online]. Available: https://www.cnas.org/publications/reports/understanding-chinas-ai-strategy
11. "Charter of the select committee on AI," White House, 2021. [Online]. Available: https://trumpwhitehouse.archives.gov/wp-content/uploads/2021/01/Charter-Select-Committee-on-AI-Jan-2021-posted.pdf

LAURA FREEMAN is the director of the intelligent systems division at the National Security Institute, Virginia Tech, Arlington, Virginia, 22203, USA. Her research interests include developing new experimental methods for characterizing emerging technologies in cyberphysical systems, artificial intelligence, and machine learning. Freeman received a Ph.D. in statistics from Virginia Tech. Contact her at laura.freeman@vt.edu.

FERAS A. BATARSEH is a research associate professor in the Bradley Department of Electrical and Computer Engineering and the Commonwealth Cyber Initiative at Virginia Tech, Arlington, Virginia, 22203, USA. His research interests lie in the areas of artificial intelligence (AI) assurance, AI for agricultural policy, cyberbiosecurity, and context-aware systems. Batarseh received a Ph.D. in computer engineering from the University of Central Florida. He is a Senior Member of IEEE. Contact him at batarseh@vt.edu.

D. RICHARD KUHN is a computer scientist in the computer security division at the National Institute of Standards and Technology, Gaithersburg, Maryland, 20899, USA. His current research interests focus on applications of combinatorial methods for assured autonomy and software verification. Kuhn received an M.S. in computer science from the University of Maryland, College Park. He is a Fellow of IEEE. Contact him at kuhn@nist.gov.

M S RAUNAK is a computer scientist in the computer security division at the National Institute of Standards and Technology, Gaithersburg, Maryland, 20899, USA. His research interests include verification and validation of "difficult to assess" software and systems to increase trust and reliability and explainable artificial intelligence. Raunak received a Ph.D. in computer science from the University of Massachusetts Amherst. He is a Member of IEEE. Contact him at raunak@ nist.gov.

RAGHU N KACKER is a senior researcher at the National Institute of Standards and Technology, Gaithersburg, Maryland, 20899, USA. His research interests include development and use of mathematical models and tool for trust and assurance of software-based systems. Kacker received a Ph.D. in statistics from the Iowa State University. He is a fellow of the American Statistical Association and the American Society for Quality. Contact him at raghu.kacker@nist.gov.