iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** XI **Month of publication:** November 2023

DOI: https://doi.org/10.22214/ijraset.2023.57178

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue XI Nov 2023- Available at www.ijraset.com

A Survey on Artificial Intelligence Assurance

Vishal Kumar[1], Shivam Kumar[2], Prashant Swamy[3], Ankit Singh[4], Md Iftekhar[5], Prof. Khusbhu Khandit[6]
Ajeenkya D Y Patil University, Charoli BK VIA Lohegaon, Pune 412105, Maharastra, India

Abstract: Artificial Intelligence (AI) algorithms are decreasingly furnishing decision-timber and functional support across multiple disciplines. AI includes a wide (and growing) library of algorithms that could be applied to different problems. One important notion for the relinquishment of AI algorithms into functional decision processes is the conception of assurance. The literature on assurance, unfortunately, conceals its issues within an involved geography of clashing approaches, driven by contradicting provocations, hypotheticals, and anticipations. Consequently, albeit a rising and new area, this handwriting provides a methodical review of exploration works that apply to AI assurance, between the times 1985 and 2021 and aims to give a structured volition to the geography. A new AI assurance description is espoused and presented, and assurance styles are varied and tabulated. also, a ten-metric scoring system is developed and introduced to estimate and compare styles. Incipiently, in this handwriting, we give foundational perceptivity, conversations, unborn directions, a roadmap, and applicable recommendations for the development and deployment of AI assurance.

I. INTRODUCTION AND SURVEY STRUCTURE

The recent rise of big data gave birth to a new pledge for AI- grounded in statistical literacy, and at this time, negative to former AI layoffs, it seems that statistical literacy-enabled AI has survived the hype, in that it has been suitable to surpass mortal- position performance in certain disciplines. analogous to any other engineering deployment, erecting AI systems requires evaluation, which may be called assurance, confirmation, verification, or another name. We address this language debate in the coming section. Defining the compass of AI assurance is worth studying, AI is presently stationed in multiple disciplines, it's soothsaying profit, guiding robots in the battleground, driving buses, recommending programs to government officers, prognosticating gravidity, and classifying guests. AI has multiple subareas similar to machine literacy, computer vision, knowledge- grounded systems, and numerous further — thus, we pose the question is it possible to give a general assurance result across all subareas and disciplines? This review sheds light on being workshop in AI assurance, provides a comprehensive overview of the state of the wisdom, and discusses patterns in AI assurance publishing. This section sets the stage for the handwriting by presenting the provocation, clear delineations, and distinctions, as well as the addition/ rejection criteria of reviewed papers.

A. Relevant Terminology and Definitions

All AI systems bear assurance; it's important to distinguish between different terms that might have been used interchangeably in the literature. We admit the following applicable terms (1) confirmation,(2) verification,(3) testing, and(4) assurance. This paper is concerned with all of the mentioned terms. The ensuing delineations are espoused in our handwriting, for clarity and to avoid nebulosity in forthcoming theoretical discussions " The process of assessing a system or element to determine whether the products of a given development phase satisfy the conditions assessed at the launch of that phase ". confirmation " The process of assessing a system or element during or at the end of the development process to determine whether it satisfies specified conditions "(Gonzalez and Barr, 2020). Another description for V&V is from the Department of Defense, as they applied testing practices to simulation systems, it states the following Verification is the " process of determining that a model perpetration directly represents the inventor's abstract descriptions and specifications ", and confirmation is the process of " determining the degree to which a model is an accurate representation "(60). According to the American Software Testing Qualification Board, testing is " the process conforming of all lifecycle conditioning, both stationary and dynamic, concerned with planning, medication, and evaluation of software products and affiliated work products to determine that they satisfy specified conditions, to demonstrate that they're fit for purpose and to descry blights ". Grounded on that (and other reviewed delineations), testing includes both confirmation and verification. Assurance this term has been infrequently applied to conventional software engineering; rather, it's used in the environment of AI and literacy algorithms. In this handwriting, grounded on previous delineations and recent AI challenges, we propose the following description for AI assurance.

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, explainable to the layman, ethical in the context of its deployment, unbiased in its learning, and fair to its user.

Our definition is by design generic and therefore applicable to all AI domains and subareas. Additionally, based on our review of a wide variety of existing definitions of assurance, it is evident that the two main AI components of interest are *the data* and *the algorithm*; accordingly, those are the two main pillars of our definition. Additionally, we highlight that the outcomes the AI enable system (intelligent system) are evaluated at the system level, where the decision or action is being taken. The remainder of this paper is focused on a review of existing AI assurance methods, and it is structured as follows: the next section presents the inclusion/exclusion criteria, "AI assurance landscape" section provides a historical perspective as well as the entire assurance landscape, "The review and scoring of methods" section includes an exhaustive list of papers relevant to AI assurance (as well as the scoring system), "Recommendations and the future of AI assurance" section presents overall insights and discussions of the survey, and lastly, "Conclusions" section presents conclusions.

B. Description of Included Articles

Articles that are included in this paper were found using the following search terms: assurance, validation, verification, and testing. Additionally, as it is well known, AI has many subareas, in this paper, the following subareas were included in the search: machine learning, data science, deep learning, reinforcement learning, genetic algorithms, agent-based systems, computer vision, natural language processing, and knowledge-based systems (expert systems). We looked for papers in conference proceedings, journals, books, and book chapters, dissertations, as well as industry white papers. The search yielded results from the year 1985 to the year 2021. Besides university libraries, multiple online repositories were searched (the most commonplace AI peer-reviewed venues). Additionally, areas of research such as data bias, data incompleteness, Fair AI, Explainable AI (XAI), and Ethical AI were used to widen the net of search. The next section presents an executive summary of the history of AI assurance.

II. AI ASSURANCE LANDSCAPE

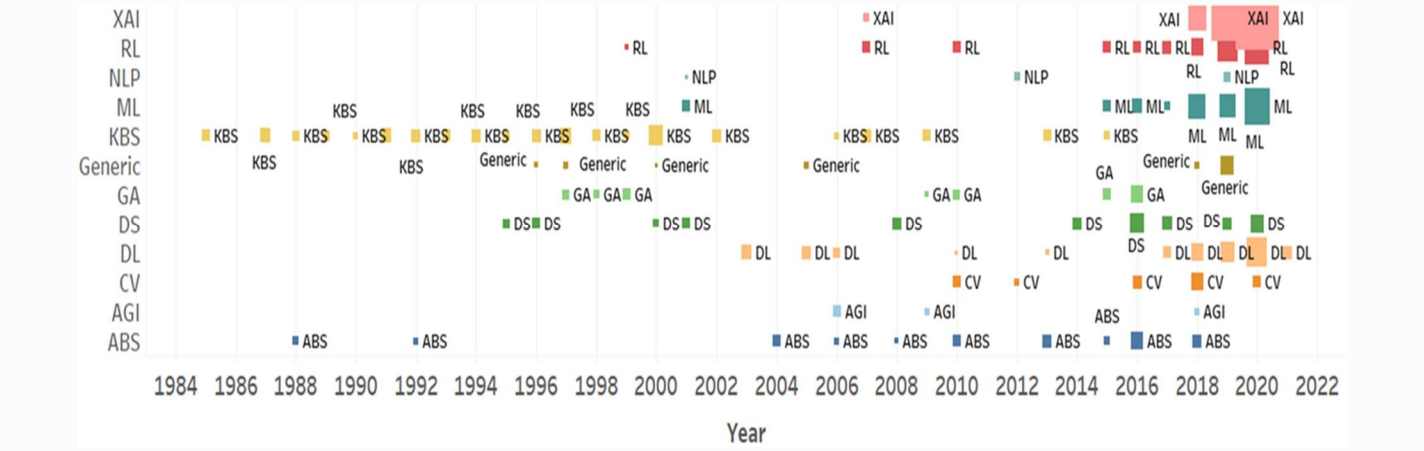
The history and current state of AI assurance is certainly a debatable matter. In this section, multiple methods are discussed, critiqued, and aggregated by AI subarea. The goal is to illuminate the need for an organized system for evaluating and presenting assurance methods; which is presented in the next sections of this manuscript.

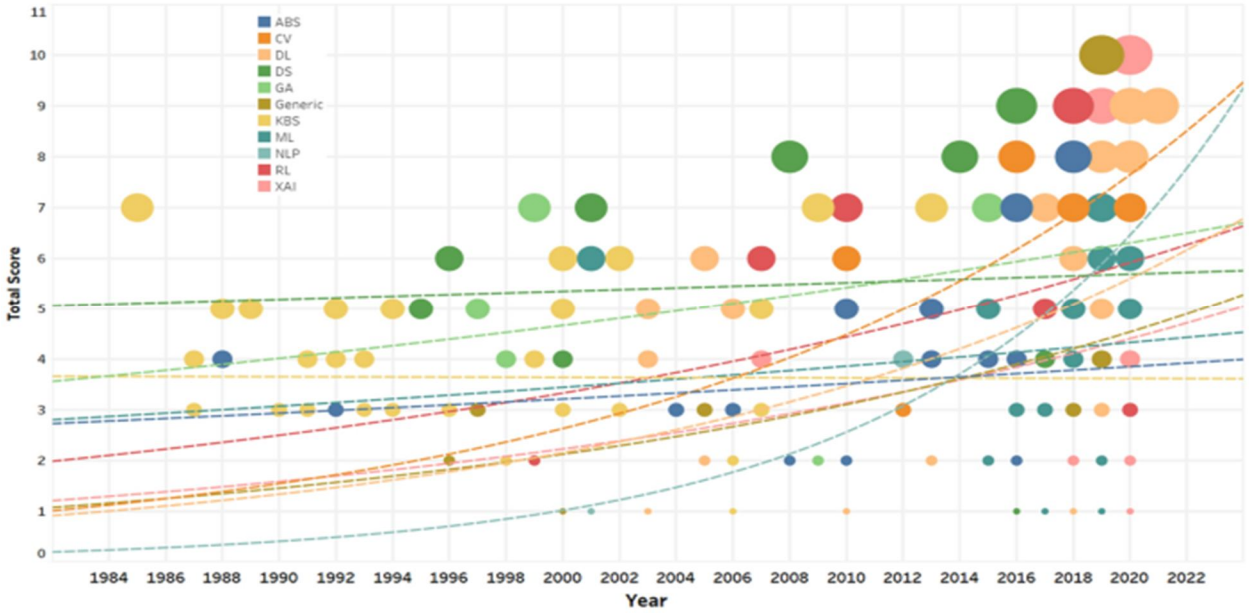
A. A historical Perspective (Analysis of the state-of-the-science)

As a starting point for AI assurance and testing, there is nowhere more suitable to begin than the Turing test [219]. In his famous manuscript: Computing Machinery and Intelligence, he introduced the imitation game, which was then popularized as the Turing test. Turing states: “The object of the game for the interrogator is to determine which of the other two is the man and which is the woman”. Based on a series of questions, the intelligent agent “learns” how to make such a distinction. If we consider the different types of intelligence, it becomes evident that different paradigms have different expectations. A genetic algorithm aims to optimize, while a classification algorithm aims to classify (choose between yes and no for instance). As Turing stated in his paper: “We are of course supposing for the present that the questions are of the kind to which an answer: Yes or No is appropriate, rather than questions such as: What do you think of Picasso?” Comparing predictions (or classifications) to actual outputs is one way of evaluating that the results of an algorithm match what the real world created.There were a dominating number of validation and verification methods in the seventies, eighties, and nineties for two forms of intelligence, knowledge-based systems (i.e., expert systems) and simulation systems (majorly for defense and military applications). One of the first times where AI turned towards data-driven methods was apparent in 1996 at the Third International Math and Science Study (TIMSS), which, focused on quality assurance in data collection (Martin and Mullis, 1996). Data from Forty-five countries were included in the analysis. In a very deliberate process, the data collectors were faced with challenges relevant to the internationalization of data. For example, data from Indonesia had errors in translation; data collection processes were different in Korea, Germany, and Kuwait than the standard process due to funding and timing issues. Such real-world issues in data collection certainly pose a challenge to the assurance of statistical learning AI that require addressing.In the 1990s, AI testing and assurance were majorly inspired by the big research archive of testing of software (i.e., within software engineering) [23]. However, a slim amount of literature explored algorithms such as genetic algorithms [104], reinforcement learning (Hailu and Sommer, 1997), and neural networks (Paladini, 1999). It was not until the 2000s that there was a serious surge in data-driven assurance and the testing of AI methods.

In the early 2000s, mostly manual methods of assurance were developed, for example, CommonKADS was a popular and commonplace method that was used to incrementally develop and test an intelligent system. Other domain-specific works were published in areas such as healthcare [27], or algorithms-specific assurance such as Crisp Clustering for k-means clustering [85].It was not until the 2010s that a spike in AI assurance for *big* data occurred. Validation of data analytics and other new areas, such as

XAI and Trustworthy AI have dominated the AI assurance field in recent years. Figure 1 illustrates that areas including XAI, computer vision, deep learning, and reinforcement learning have had a recent spike in assurance methods; and the trend is expected to be increasingly on the rise (as shown in Fig. 2). The figure also illustrates that knowledge-based systems were the focus until the early nineties, and shows a shift towards the statistical learning based subareas in the 2010s. A version of the dashboard is available in a public repository (with instructions on how to run it): https://github.com/ferasbatarseh/AI-Assurance-Review.





The p-values for the trend lines presented in Fig. 2 are as follows: Data Science (DS): 0.87, Genetic Algorithms (GA): 0.50, Reinforcement Learning (RL): 0.15, Knowledge-Based Systems (KBS): 0.97, Computer Vision (CV): 0.22, Natural Language Processing (NLP): 0.17, Generic AI: 0.95, Agent-Based Systems (ABS): 0.33, Machine Learning (ML): 0.72, Deep Learning (DL): 0.37, and XAI: 0.44.It is undeniable that there is a rise in the research of AI, and especially in the area of assurance. The next section "The state of AI assurance" provides further details on the state-of-the-art, and "The review and scoring of methods" section presents an exhaustive review of all AI assurance methods found under the predefined search criteria.

B. The State of AI Assurance

This section introduces some milestone methods and discussion in AI assurance. Many of the discussed works rely on standard software validation and verification methods. Such methods are inadequate for AI systems, because they have a dimension of intelligence, learning, and re-learning, as well as adaptability to certain contexts.

Therefore, errors in AI system "may manifest themselves because of autonomous changes" [211], and among other scenarios would require extensive assurance. For instance, in expert systems, the inference engine component creates rules and new logic based on forward and backward propagation [20]. Such processes require extensive assurance of the process as well as the outcome rules. Alternatively, for other AI areas such as neural networks, while propagation is used, taxonomic evaluations and adversarial targeting are more critical to their assurance [145]. For other subareas such as machine learning, the structure of data, data collection decisions, and other data-relevant properties need step-wise assurance to evaluate the resulted predictions and forecasts. For instance, several types of bias can occur in any phase of the data science lifecycle or while extracting outcomes. Bias can begin during data collection, data wrangling, modeling, or any other phase. Biases and variances which arise in the data are independent of the sample size or statistical significance, and they can directly affect the *context* or the results or the model. Other issues such as incompleteness, data skewness, or lack of structure have a negative influence on the quality of outcomes of any AI model and require data assurance [117].While the historic majority of methods for knowledge-based systems and expert systems (as well as neural networks) aimed at finding generic solutions for their assurance [21, 218], and [166], other "more recent" methods were focused on one AI subarea and one domain. For instance, in Mason et al. [142, 144], assurance was applied to reinforcement learning methods for safety–critical systems. Prentzas et al. [174] presented an assurance method for machine learning as its applied to stroke predictions, similar to Pawar's et al.'s [167] XAI for healthcare framework. Pepe et al. [169], and Chittajallu et al.'s [42] developed a method for surgery video detection methods. Moreover, domains such as law and society would generally benefit from AI subareas such as natural language processing for analyzing legal contracts [135], but also require assurance.Another major aspect (for most domains) that was evident in the papers reviewed was the need for explainability (i.e. XAI) of the learning algorithm, defined as: *to identify how the outcomes were arrived at* (transforming the black-box to a white-box) [193]. Few papers without substantial formal methods were found for Fair AI, Safe AI [68], Transparent AI [1], or Trustworthy AI [6], but XAI [83] has been central (as the previous figures in this paper also suggest). For instance, in Lee et al. [121], layer-wise relevance propagation was introduced to obtain the effects of every neural layer and each neuron on the outcome of the algorithm. Those observations are then presented for better understanding of the model and its inner workings. Additionally, Arrieta et al. [16] presented a model for XAI that is tailored for road traffic forecasting, and Guo [82] presented the same, albeit for 5G and wireless networks [200]. Similarly, Kuppa and Le-Khac [118] presented a method focused on Cyber Security using gradient maps and bots. Go and Lee [76] presented an AI assurance method for trustworthiness of security systems. Lastly, Guo [82] developed a framework for 6G testing using deep neural networks.Multi-agent AI is another aspect that requires a specific kind of assurance, by validating every agent, and verifying the integration of agents [163]. The challenges of AI algorithms and their assurance is evident and consistent across many of the manuscripts, such as in Janssen and Kuk's [100] study of the limitations of AI for government, on the other hand, Batarseh et al. [22] presented multiple methods for applying data science at government (with assurance using knowledge-based systems). Assurance is especially difficult when it comes to being performed in real time, timeliness in critical systems, and other defense-relevant environments is very important [54, 105, 124], and (Laat, 2017). Other less "time-constrained" activities such as decisions at organizations [186] and time series decision support systems could utilize slower methods such as genetic algorithms [214], but they require a different take on assurance. The authors suggested that "by no means we have a definitive answer, what we do here is intended to be suggestive" [214] when addressing the validation part of their work. A recent publication by Raji et al. [180] shows a study from the Google team claiming that they are "aiming to close the accountability gap of AI" using an internal audit system (at Google). IBM research also proposed few solutions to manage the bias of AI services [202, 225]. As expected, the relevance and success of assurance methods varied, and so we developed a scoring system to evaluate existing methods. We were able to identify 200 + relevant manuscripts with methods. The next section presents the exhaustive list of the works presented in this section in addition to multiple others with our derived scores.

III. THE REVIEW AND SCORING OF ASSURANCE METHODS

The scoring of each AI assurance method/paper was based on the sum of the score of ten metrics. The objective of the metrics is to provide readers with a meaningful strategy for sorting through the vast literature on AI assurance. The scoring metric is based on the authors' review of what makes a useful reference paper for AI assurance. Each elemental metric is allocated one point, and each method is either given that point or not (0 or 1), as follows:

1) *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. specific) on the inner workings of AI systems.
2) *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).

3) *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.
4) *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.
5) *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.
6) *Declared Success:* Whether the authors declared success of their method in reaching an *assured* AI system (1) or not (0).
7) Mentioned limitations: whether there are obvious method limitations (0) or not (1).
8) *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.

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.

Table 1 presents the methods reviewed, along with their first author's last name, publishing venue, AI subarea, as well as the score (sum of ten metrics).Other aspects such as domain of application were missing from many papers and inconsistent, therefore, we didn't include them in the table. Additionally, we considered *citations per paper.* However, the data on citations (for a 250+ papers study) were incomplete and difficult to find in many cases. For many of the papers, we did not have information on how many times they were cited, because many publishers failed to index their papers across consistent venues (e.g., Scopus, MedLine, Web of Science, and others). Additionally, the issue of *self-citation* is in some cases considered in scoring but in other cases is not. Due to these citation inconsistencies (which are believed to be a challenge that reaches all areas of science), we deemed that using citations would provide more questions than answers than our subject matter expert based metrics.Appendix presents a list of all reviewed manuscripts and their detailed scores (for the ten metrics) by ranking category; ten columns matching the presented scoring method, as follows: AI subarea: AI.s; Relevance: R; Method: M; Results: Rs; Dataset: Ds; Size: Sz; Success: Sc; Limitations: L; General: G; Application: A; and Comparison: C. The papers, data, dashboard, and lists are on a public GitHub repository: https://github.com/ferasbatarseh/AI-Assurance-Review.In 2018, AI papers accounted for 3% of all peer reviewed papers published worldwide [181]. The share of AI papers has grown three-fold over twenty years. Moreover, between 2010 and 2019, the total number of AI papers on arXiv increased over 20-fold [181]. As of 2019, machine learning papers have increased most dramatically, followed by computer vision and pattern recognition. While machine learning was the most active research areas in AI, its subarea, DL have become increasing popularly in the past few years. According to GitHub, TensorFlow is the most popular free and open-source software library for AI. TensorFlow is a corporate-backed research framework, and it has been shown that, in recent years, there's noticeable trend of the emergence of such corporate-backed research frameworks. Since 2005, attendances at large AI conferences have grown significantly, NeurIPS and ICML (being the two fastest growing conferences have over eight-fold increase. Attendances at small AI conferences have also grown over 15-fold starting from 2014, and the increase is highly related to the emergence of deep and reinforcement learning [181]. As the field of AI continues to grow, assurance of AI has become a more important and timely topic.

IV. RECOMMENDATIONS AND THE FUTURE OF AI ASSURANCE

A. The Need for AI Assurance

The emergence of complex, opaque, and invisible algorithms that learn from data motivated a variety of investigations, including: algorithm awareness, clarity, variance, and bias [94]. Algorithmic bias for instance, whether it occurs in an unintentional or intentional manner, is found to severely limit the performance of an AI model. Given AI systems provide recommendations based on data, users' faith in that the recommended outcomes are trustworthy, fair, and not biased is another critical challenge for AI assurance. Applications of AI such as facial recognition using deep learning have become commonplace. Deep learning models are often exposed to adversarial inputs (such as deep-fakes), thus limiting their adoption and increasing their threat [145]. Unlike conventional software, aspects such as explainability (unveiling the blackbox of AI models) dictate how assurance is performed and what is needed to accomplish it.

Unfortunately however, similar to the software engineering community's experience with testing, ensuring a valid and verified system is often an afterthought. Some of the classical engineering approaches would prove useful to the AI assurance community, for instance, performing testing in an incremental manner, involving users, and allocating time and budget specifically to testing, are some main lessons that ought to be considered.

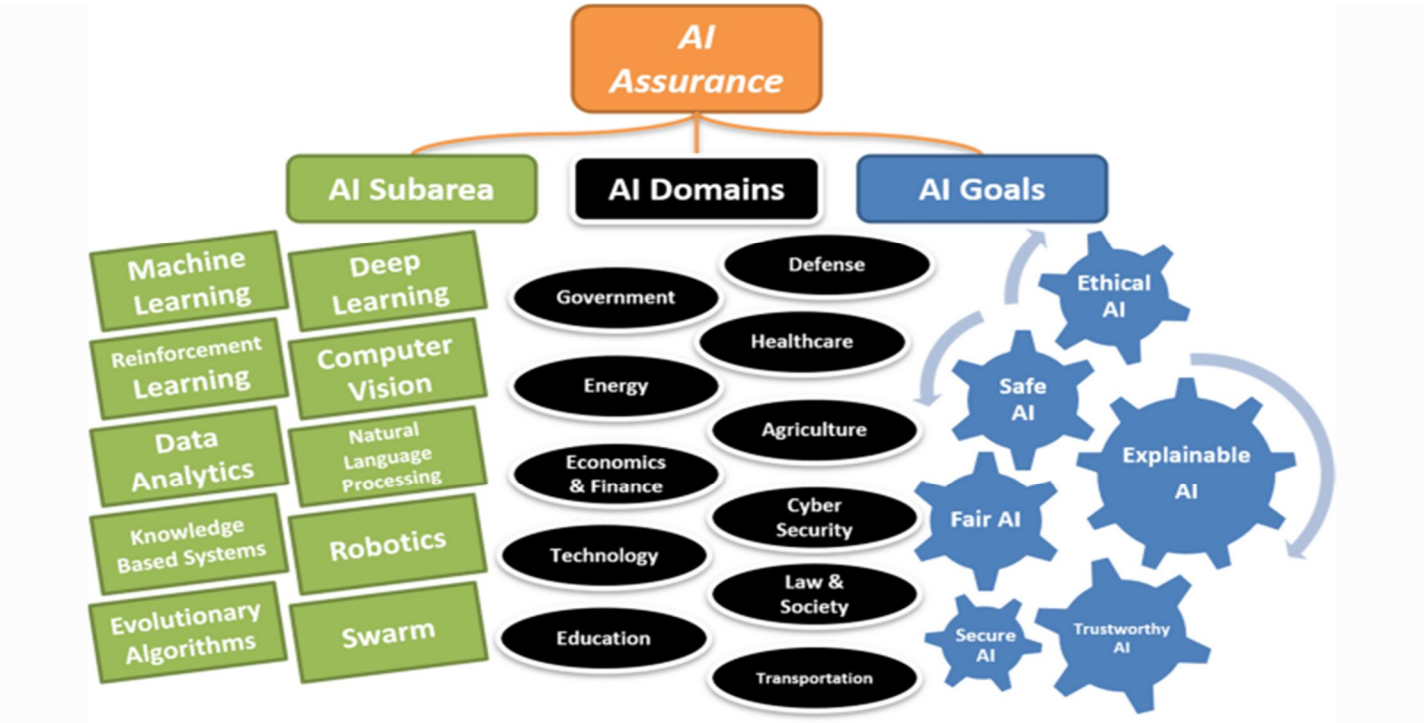
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue XI Nov 2023- Available at www.ijraset.com

A worthy recent trend that might aid majorly in assurance is using AI for testing AI (i.e., deploying intelligence methods for the testing and assurance of AI methods). Additionally, from a user's perspective, recent growing questions in research that are relevant to assurance pose the following concerns: how is learning performed inside the blackbox? How is the algorithm creating its outcomes? Which dependent variables are the most influential? Is the AI algorithm dependable, safe, secure, and ethical? Besides all the previously mentioned assurance aspects, we deem the following foundational concepts as highly connected, worthy of considering by developers and AI engineers, and essential to all forms of AI assurance: (1) Context: refers to the scope of the system, which could be associated with a timeframe, a geographical area, specific set of users, and any other system environmental specifications (2) Correlation: the amount of relevance between the variables, this is usually part of exploratory analysis, however, it is key to understand which dependent variables are correlated and which ones are not, (3) Causation: the study of cause and effect; i.e., which variables directly cause the outcome to change (increase or decrease) in any fashion, (4) Distribution: whether a normal distribution is assumed or not. Data distribution of the inputted dependent variables can dictate which models are best suited for the problem at hand, and (5) Attribution: aims at allocating the variables in the dataset that have the strongest influence on the outcomes of the AI algorithm.Providing a scoring system to evaluate existing methods provides support to scholars in evaluating the field, avoiding future mistakes, and creating a system where AI scientific methods are measured and evaluated by others, a practice that is becoming increasingly rare in scientific arenas. More importantly, practitioners –in most cases– find it difficult to identify the best method for assurance relevant to their domain and subarea. We anticipate that this comprehensive review will help in that regard as well. As part of AI assurance, ethical outcomes should be evaluated, while ethical considerations might differ from one context to another, it is evident that requiring outcomes to be ethical, fair, secure, and safe necessitates the involvement of humans, and in most cases, experts from other domains. That notion qualifies AI assurance as a multidisciplinary area of investigation

.

V. FUTURE COMPONENTS OF AI ASSURANCE RESEARCH

In some AI subareas, there are known issues to be tackled by AI assurance, such as deep learning's sensitivity to adversarial attacks, as well as overfitting and underfitting issues in machine learning. Based on that and on the papers reviewed in this survey, it is evident that AI assurance is a necessary pursuit, but a difficult and multi-faceted area to address. However, previous experiences, successes, and failures can point us to what would work well and what is worth pursuing. Accordingly, we suggest performing and developing AI assurance by (1) domain, by (2) AI sub area, and by (3) AI goal; as a theoretical roadmap, similar to what is shown in Fig. 3.

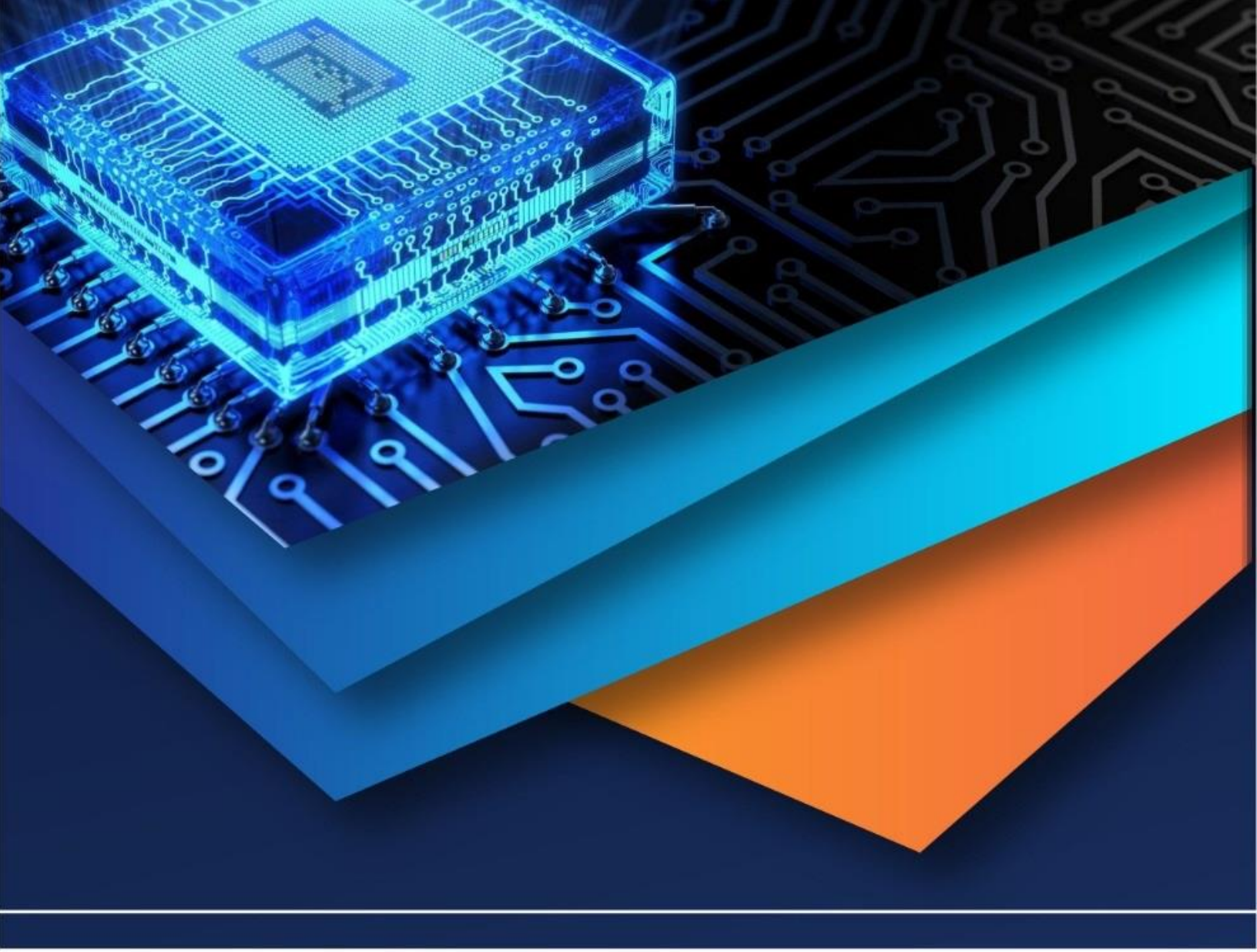


Three-dimensional AI assurance by subarea, domain, and goal
Fig 3

In some cases, such as in unsupervised learning techniques, it is difficult to know what to validate or assure [86]. In such cases, the outcome is not predefined (contrary to supervised learning). Genetic algorithms and reinforcement learning have the same issue, and so in such cases, feature selection, data bias, and other data-relevant validation measures, as well as hypothesis generation and testing become more important. Additionally, different domains require different tradeoffs; trustworthiness for instance is more important when it comes to using AI in healthcare versus when its being used for revenue estimates at a private sector firm; also, AI safety is more critical in defense systems than in systems built for education or energy application.Other surveys presented a review of AI validation and verification [71] and [21], however, none was found that covered the three dimensional structure presented (by subarea, goal, and domain) like this review.

VI. CONCLUSIONS

In AI assurance, there are other philosophical questions that are also very relevant, such as what is a valid system? What is a trustworthy outcome? When to stop testing or model learning? When to claim victory on AI safety? When to allow human intervention (and when not to)? And many other similar questions that require close attention and evaluation by the research community. The most successful methods presented in literature (scored as 8, 9, or 10), are the ones that were specific to an AI *subarea* and *goal;* additionally, ones that had done extensive theoretical and hands-on experimentation. Accordingly, we propose the following five considerations as they were evident in existing successful works when defining or applying new AI assurance methods: (1) *Data quality*: similar to assuring the outcomes, assuring the dataset and its quality mitigates issues that would eventually prevail in the AI algorithm. (2) *Specificity*: as this review concluded, the assurance methods ought to be designed to one goal and subarea of AI. (3) *Addressing invisible issues*: AI engineers should carry out assurance in a procedural manner, not as an afterthought or a process that is performed only in cases of the presence of visible issues. (4) *Automated assurance*: using manual methods for assurance would in many cases defeat the purpose. It is difficult to evaluate the validity of the assurance method itself, hence, automating the assurance process can—if done with best practices in mind– minimize error rates due to human interference. (5) *The user*: involving the user in an incremental manner is critical in expert-relevant (non-engineering) domains such as healthcare, education, economics, and other areas. Explainability is a relative and subjective matter; hence, users of the AI system can help in defining how explainability ought to be presented.Based on all discussions presented, we assert it will be beneficial to have multi-disciplinary collaborations in the field of AI assurance. The growth of the field might need not only computer scientists and engineers to develop advanced algorithms, but also economists, physicians, biologists, lawyers, cognitive scientists, and other domain experts to unveil AI deployments to their domains, create a data-driven culture within their organizations, and ultimately enable the wide-scale adoption of assured AI systems.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY