Batarseh *et al. J Big Data* (2021) 8:60
https://doi.org/10.1186/s40537-021-00445-7

Journal of Big Data

SURVEY PAPER Open Access

A survey on artificial intelligence assurance

Feras A. Batarseh[1*], Laura Freeman[2] and Chih-Hao Huang[3]

*Correspondence:
batarseh@vt.edu
[1] The Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University (Virginia Tech), Arlington, VA 22203, USA
Full list of author information is available at the end of the article

Abstract

Artificial Intelligence (AI) algorithms are increasingly providing decision making and operational support across multiple domains. AI includes a wide (and growing) library of algorithms that could be applied for different problems. One important notion for the adoption of AI algorithms into operational decision processes is the concept of assurance. The literature on assurance, unfortunately, conceals its outcomes within a tangled landscape of conflicting approaches, driven by contradicting motivations, assumptions, and intuitions. Accordingly, albeit a rising and novel area, this manuscript provides a systematic review of research works that are relevant to AI assurance, between years 1985 and 2021, and aims to provide a structured alternative to the landscape. A new AI assurance definition is adopted and presented, and assurance methods are contrasted and tabulated. Additionally, a ten-metric scoring system is developed and introduced to evaluate and compare existing methods. Lastly, in this manuscript, we provide foundational insights, discussions, future directions, a roadmap, and applicable recommendations for the development and deployment of AI assurance.

Keywords: AI assurance, Data Engineering, Explainable AI (XAI), Validation and verification

Introduction and survey structure

The recent rise of big data gave birth to a new promise for AI based in statistical learning, and at this time, contrary to previous AI winters, it seems that statistical learning enabled AI has survived the hype, in that it has been able to surpass human-level performance in certain domains. Similar to any other engineering deployment, building AI systems requires evaluation, which may be called assurance, validation, verification or another name. We address this terminology debate in the next section.

Defining the scope of AI assurance is worth studying, AI is currently deployed at multiple domains, it is forecasting revenue, guiding robots in the battlefield, driving cars, recommending policies to government officials, predicting pregnancies, and classifying customers. AI has multiple subareas such as machine learning, computer vision, knowledge-based systems, and many more—therefore, we pose the question: is it possible to provide a generic assurance solution across all subareas and domains? This review sheds light on existing works in AI assurance, provides a comprehensive overview of the *state-of-the-science*, and discusses patterns in AI assurance publishing. This section sets that

stage for the manuscript by presenting the motivation, clear definitions and distinctions, as well as the inclusion/exclusion criteria of reviewed articles.

Relevant terminology and definitions

All AI systems require assurance; it is important to distinguish between different terms that might have been used interchangeably in literature. We acknowledge the following relevant terms: (1) validation, (2) verification, (3) testing, and (4) assurance. This paper is concerned with all of the mentioned terms. The following definitions are adopted in our manuscript, for the purposes of clarity and to avoid ambiguity in upcoming theoretical discussions:

Verification: "The process of evaluating a system or component to determine whether the products of a given development phase satisfy the conditions imposed at the start of that phase". Validation: "The process of evaluating a system or component during or at the end of the development process to determine whether it satisfies specified requirements" (Gonzalez and Barr, 2020). Another definition 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 implementation accurately represents the developer's conceptual descriptions and specifications", and Validation is the process of "determining the degree to which a model is an accurate representation" [60].

Testing: according to the American Software testing Qualification Board, testing is "the process consisting of all lifecycle activities, both static and dynamic, concerned with planning, preparation and evaluation of software products and related work products to determine that they satisfy specified requirements, to demonstrate that they are fit for purpose and to detect defects". Based on that (and other reviewed definitions), testing includes both validation and verification.

Assurance: this term has been rarely applied to conventional software engineering; rather, it is used in the context of AI and learning algorithms. In this manuscript, based on prior definitions and recent AI challenges, we propose the following definition 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 and explainable to a layman, ethical in the context of its deployment, unbiased in its learning, and fair to its users**.

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 remaining 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.

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 year 1985 to 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.

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 next sections of this manuscript.

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,

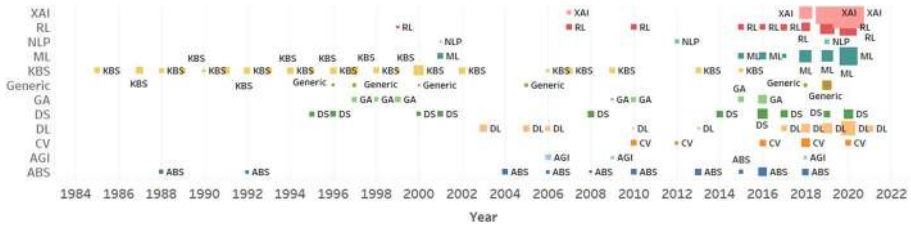

Fig. 1 A history of AI assurance by year and subarea

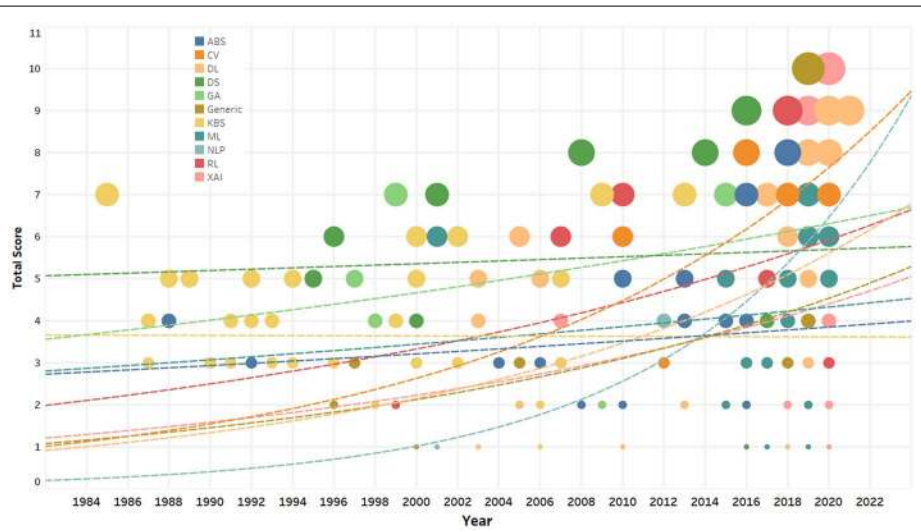

Fig. 2 Past and future (using trend lines) of AI assurance (by AI subarea)

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.

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.

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:

I. 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.
II. 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).
III. 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.
IV. 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.
V. 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.
VI. Declared success: whether the authors declared success of their method in reaching an *assured* AI system (1) or not (0).
VII. Mentioned limitations: whether there are obvious method limitations (0) or not (1).
VIII. 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.
IX. 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

Table 1 Reviewed methods and their total scores

Year	First author's last name and citation	Publishing venue	AI subarea	Total score
2020	D'Alterio [50]	FUZZ-IEEE	XAI	10
2019	Tao [208]	IEEE Access	Generic	10
2020	Anderson [11]	ACM TIIS	RL	9
2020	Birkenbihl [29]	EPMA	ML	9
2020	Checco [39]	JAIR	DS	9
2020	Chen [40]	IEEE Access	XAI	9
2020	Cluzeau [43]	EASA	DL	9
2019	Kaur [109]	WAINA	XAI	9
2020	Kulkarni [117]	Academic Press	DS	9
2020	Kuppa [118]	IEEE IJCNN	XAI	9
2020	Kuzlu [120]	IEEE Access	XAI	9
2021	Massoli [145]	CVIU	DL	9
2020	Spinner [201]	IEEE TVCG	XAI	9
2016	Veeramachaneni [226]	IEEE HPSC	DS	9
2018	Wei [230]	AS	RL	9
2020	Winkel [236]	EJR	RL	9
2014	Ali [8]	GISci	DS	8
2018	Alves [9]	NASA ARIAS	ABS	8
2019	Batarseh [24]	EDML	DS	8
2016	Gao [71]	SEKE	DS	8
2020	Gardiner [72]	Nature Sci Rep	ML	8
2016	Gulshan [81]	JAMA	CV	8
2020	Guo [82]	IEEE ICCVW	XAI	8
2020	Han [87]	IET JoE	XAI	8
2016	Heaney [93]	OD	GA	8
2019	Huber [97]	KI AAI	RL	8
2019	Keneni [112]	IEEE Access	XAI	8
2020	Kohlbrenner [116]	IEEE IJCNN	XAI	8
2019	Maloca [134]	PLoS ONE	DL	8
2020	Malolan [136]	IEEE ICICT	XAI	8
2020	Payrovnaziri [168]	JAMIA	ML	8
2008	Peppler [170]	OASJ	DS	8
2020	Sequeira [196]	SciDir AI	RL	8
2020	Sivamani [199]	IEEE LCS	DL	8
2020	Tan [207]	IEEE IJCNN	XAI	8
2020	Tao [209]	IEEE CoG	XAI	8
2020	Welch [231]	PhysMedBiol	DL	8
2020	Xiao [239]	IS	DL	8
2016	Aitken [6]	UC	ABS	7
2019	Barredo-Arrieta [19]	IEEE ITSC	XAI	7
2013	Batarseh [20]	IEEE TSMCS	KBS	7
2001	Berndt [27]	COMP	DS	7
2010	Bone [30]	CEUS	RL	7
2016	Celis [38]	PrePrint	ML	7
2019	Chittajallu [42]	IEEE ISBI	XAI	7
2018	Elsayed [67]	NIPS	CV	7
2019	Ferreyra [69]	FUZZ-IEEE	XAI	7
2006	Forster [70]	Uni of South Africa	AGI	7
1985	Ginsberg [74]	IJCAI	KBS	7

Table 1 (continued)

Year	First author's last name and citation	Publishing venue	AI subarea	Total score
2018	Go [76]	ACM CCS	DL	7
2020	Halliwell [86]	PrePrint	DL	7
2015	He [90]	MPE	GA	7
2020	Heuer [94]	ACM UMAP	ML	7
2016	Jiang [102]	PMLR	RL	7
2020	Kaur [111]	AINA	XAI	7
2016	Kianifar [113]	SC	GA	7
2019	Lee [121]	IEEE ICTC	XAI	7
2017	Liang [126]	MILCOM	DS	7
2020	Mackowiak [132]	PrePrint	CV	7
2018	Mason [143]	AHIM	RL	7
2018	Murray [157]	FUZZ-IEEE	XAI	7
2019	Naqa [66]	MedPhys	ML	7
2019	Prentzas [174]	IEEE BIBE	XAI	7
2018	Pynadath [177]	Springer HCIS	ML	7
2020	Ragot [179]	CHI	ML	7
2020	Rotman [184]	PrePrint	RL	7
2015	Rovcanin [185]	WN	RL	7
2020	Sarathy [188]	IEEE SISY	XAI	7
2018	Stock [203]	ECCV	CV	7
2009	Tadj [206]	SCI	KBS	7
1999	Thomas [214]	AAAI	GA	7
2020	Uslu [220]	AINA	XAI	7
2018	Xu [240]	PrePrint	DL	7
2019	Bellamy [26]	IBM JRD	XAI	6
2019	Beyret [28]	IEEE IROS	RL	6
2018	Cao [35]	JAIHC	ML	6
2020	Cruz [47]	PrePrint	RL	6
2001	Halkidi [85]	JIIS	ML	6
2020	He [91, 92]	PrePrint	RL	6
2020	Islam [98]	IEEE TFS	XAI	6
2005	Liu [128]	AI2005	DL	6
2019	Madumal [133]	PrePrint	RL	6
1996	Martin [138]	ERIC	DS	6
2007	Martín-Guerrero [141]	AJCAI	RL	6
2000	Mosqueira-Rey [155]	ESA	KBS	6
2020	Mynuddin [160]	IETITS	RL	6
2020	Puiutta [175]	CD-MAKE	RL	6
2018	Ruan [187]	IJCAI	DL	6
2019	Schlegel [193]	IEEE ICCVW	XAI	6
2020	Toreini [216, 217]	ACM FAT	ML	6
2020	Toreini [216, 217]	PrePrint	ML	6
2019	Vabalas [222]	PLoS ONE	ML	6
2010	Winkler [237]	IEEE SUTC	CV	6
2002	Wu [238]	IJHCS	KBS	6
2019	Zhu [246]	ACM PLDI	RL	6
1992	Andert [12]	IJM	KBS	5
2018	Antunes [14]	IEEE DSN-W	ML	5
1989	Becker [25]	NASA	KBS	5

Table 1 (continued)

Year	First author's last name and citation	Publishing venue	AI subarea	Total score
2019	Chen [41]	CS	RL	5
2019	Cruz [46]	AI 2019 AAI	RL	5
2020	Diallo [57]	IEEE ACSOS-C	XAI	5
2010	Dong [62]	IEEE ICWIIAT	GA	5
2019	Dupuis [64]	UoG	XAI	5
2015	Goodfellow [79]	PrePrint	ML	5
2020	Guo [82]	IEEE CM	XAI	5
2020	Haverinen [89]	Uni of Jyväskylä	XAI	5
1997	Jones [104]	JMB	GA	5
2019	Joo [106]	IEEE CoG	RL	5
2020	Katell [107]	ACM FAT	XAI	5
2007	Knauf [115]	IEEE TSMC	KBS	5
1995	Lockwood [130]	AES	KBS	5
2000	Marcos [137]	IEE Proc	KBS	5
2017	Mason [144]	WhiteRose	RL	5
1988	Morell [154]	IEA/AIE	KBS	5
2020	Murray [158]	IEEE TETCI	XAI	5
2010	Niazi [162]	SpringSim	ABS	5
2000	Onoyama [166]	JETAI	KBS	5
2019	Ren [182]	PrePrint	DL	5
2013	Sargent [189]	JoS	ABS	5
2003	Schumann [195]	EANN	DL	5
1995	Singer [198]	POQ	DS	5
2019	Srivastava [202]	AAAI AIES	NLP	5
2006	Taylor [211]	Springer	DL	5
2020	Taylor [213]	IEEE CVPRW	XAI	5
2020	Tjoa [215]	IEEE TNNLS	ML	5
2020	Uslu [221]	BWCCA	XAI	5
2020	Varshney [225]	IEEE CISS	ML	5
2018	Volz [228]	IEEE CIG	XAI	5
2020	Wieringa [234]	ACM FAT	XAI	5
2020	Wing [235]	PrePrint	ML	5
2019	Yoon [242]	IEEE ICCVW	XAI	5
2019	Zhou [245]	IJCAI XAI	ML	5
1994	Zlatareva [249]	ESA	KBS	5
2018	AI Now (Algorithmic Accountability Policy Tooklit) [7]	AI Now	XAI	4
2015	Arifin [15]	Springer	ABS	4
2015	Batarseh [21]	AIR	KBS	4
2007	Brancovici [31]	IEEE CEC	XAI	4
1987	Castore [37]	NASA STI	KBS	4
2013	Cohen [45]	EternalS	NLP	4
2020	Das [51]	PrePrint	XAI	4
2013	David [52]	UCS	ABS	4
2018	Došilović [63]	MIPRO	ML	4
2000	Edwards [65]	Oxford	DS	4
2018	EY (Assurance in the Age of AI) [17]	EY	ML	4
2019	Guidotti [80]	ACM CS	XAI	4
2018	Jilk [103]	PrePrint	ABS	4
2017	Leibovici [123]	ISPRS Int J. Geo-Inf	DS	4

Table 1 (continued)

Year	First author's last name and citation	Publishing venue	AI subarea	Total score
2020	Li [125]	IEEE TKDE	XAI	4
2019	Mehrabi [146]	PrePrint	ML	4
2019	Meskauskas [150]	FUZZ-IEEE	XAI	4
1998	Miller [151]	MS	GA	4
2019	Nassar [161]	WIREs DMKD	XAI	4
1992	Preece [173]	ESA	KBS	4
2019	Qiu [178]	AS	Generic	4
1984	Sargent [190]	IEEE WSC	ABS	4
2003	Taylor [212]	SPIE	DL	4
1999	Tsai [218]	IEEE TKDE	KBS	4
1991	Vinze [227]	IM	KBS	4
2019	Wang [229]	ACM CHI	XAI	4
1993	Wells [232]	AAAI	KBS	4
2018	Zhu [247]	IEEE CIG	XAI	4
1998	Zlatareva [248]	DBLP	KBS	4
2018	Abdollahi [1]	Springer	ML	3
1997	Abel [2]	FLAIRS Conference	KBS	3
2018	Adadi [4]	IEEE Access	XAI	3
2018	Agarwal [5]	PrePrint	Generic	3
2016	Amodei [10]	PrePrint	ML	3
2019	Breck [32]	SysML	ML	3
1996	Carley [36]	CASOS	KBS	3
2000	Coenen [44]	CUP	KBS	3
1987	Culbert [48]	NASA SOAR	KBS	3
2020	Dağlarli [49]	ADL	XAI	3
1992	Davis [53]	RAND	ABS	3
2020	Dodge [61]	ExSS-ATEC	XAI	3
2018	Everitt [68]	IJCAI	AGI	3
1991	Gilstrap [73]	TI	KBS	3
2019	Glomsrud [75]	ISSAV	XAI	3
1996	Gonzalez [78]	EAAI	KBS	3
1997	Harmelen [88]	EUROVAV	KBS	3
2019	He [91, 92]	PrePrint	DL	3
2020	Heuillet [95]	PrePrint	RL	3
2009	Hibbard [96]	AGI	AGI	3
2019	Israelsen [99]	ACM CSUR	Generic	3
2019	Jha [101]	NeurIPS	DL	3
2002	Knauf [114]	IEEE TSMC	KBS	3
2017	de Laat [54]	PhilosTechnol	ML	3
1994	Lee [122]	IEEE TSMC	KBS	3
2004	Liu [127]	IEEE MLC	ABS	3
1997	Lowry [131]	ISMIS	Generic	3
2012	Martinez-Balleste [139]	IEEE SIPC	CV	3
2020	Martinez-Fernandez [140]	PrePrint	XAI	3
2017	Mason [142]	DCAART	RL	3
1993	Mengshoel [148]	IEEE exp	KBS	3
2005	Menzies [149]	AC	Generic	3
2007	Min [152]	WSC	KBS	3
1997	Murrell [159]	DSS	KBS	3

Table 1 (continued)

Year	First author's last name and citation	Publishing venue	AI subarea	Total score
1987	O'Keefe [164]	IEEE exp	KBS	3
2020	Putzer [176]	PrePrint	XAI	3
1991	De Raedt [55]	JWS	KBS	3
2020	Raji [180]	ACM FAT	XAI	3
2004	Sargent [191]	IEEE WSC	ABS	3
1990	Suen [204]	ESA	KBS	3
2019	Sun [205]	IEEE VTC	XAI	3
2006	Yilmaz [241]	CMOT	ABS	3
1997	Zaidi [243]	Automatica	KBS	3
1996	Abel [3]	FLAIRS Conference	KBS	2
2016	Aitken [6]	PrePrint	ABS	2
1998	Antoniou [13]	AI Magazine	KBS	2
2019	Arrieta [16]	SciDir IF	XAI	2
2018	Bride [34]	ICFEM	XAI	2
2020	Dghaym [56]	AU SSAV	XAI	2
2015	Dobson [59]	JCLS	ML	2
2018	Hagras [83]	IEEE Comp	XAI	2
1999	Hailu [84]	IEEE SMC	RL	2
2020	He [91, 92]	IEEE IRCE	XAI	2
2016	Janssen [100]	GIQ	DS	2
2020	Kaur [110]	NBiS	XAI	2
2008	Liu [129]	IEEE SSSC	ABS	2
2006	Min [153]	ICMLC	KBS	2
2019	Mueller [156]	PrePrint	XAI	2
1996	Nourani [163]	ACM SIGSOFT	Generic	2
2020	Pawar [167]	IEEE CyberSA	XAI	2
2009	Pèpe [169]	JCG	GA	2
2013	Pitchforth [171]	ESA	DL	2
2017	Protiviti (Validation of Machine Learning Models) [223]	Protiviti	ML	2
2010	Sargent [192]	WSC	ABS	2
2019	Spada [200]	AIAI	XAI	2
2005	Taylor [210]	IEEE IJCNN	DL	2
2016	Zeigler [244]	JDMS	ABS	2
2001	Barr [18]	ACL	NLP	1
2020	Brennen [33]	ACM CHI EA	XAI	1
2006	Dibie-Barthélemy [58]	KBS	KBS	1
2020	European Commission (A European Approach to Excellence and Trust) [165]	European Commission	XAI	1
2000	Gonzalez [77]	JETAI	Generic	1
2018	Kaul [108]	ACM AIES	ML	1
2003	Kurd [119]	SAFECOMP	DL	1
2017	Lepri [124]	PhilosTechnology	ML	1
2018	Mehri [147]	ACM ARES	DL	1
2019	Pocius [172]	AAAI-19	RL	1
2019	Rossi [183]	JIA	XAI	1
2010	Schumann [194]	NASA SCI	DL	1
2018	Sileno [197]	PrePrint	XAI	1
2019	Varshney [224]	ACM XRDS	ML	1

Table 1 (continued)

Year	First author's last name and citation	Publishing venue	AI subarea	Total score
2016	Wickramage [233]	FTC	DS	1

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/feras batarseh/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.

Recommendations and the future of AI assurance

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 deepfakes), 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. 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.

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

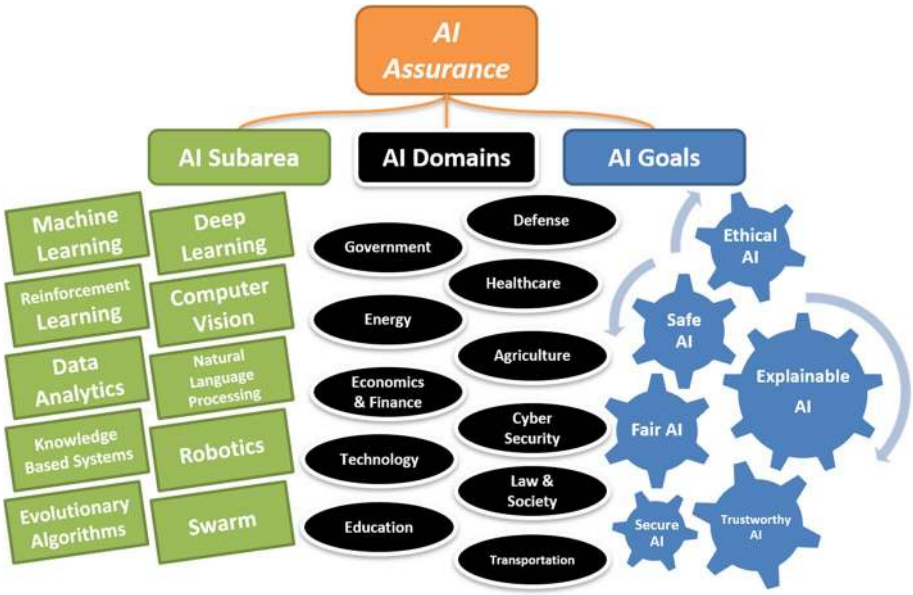

Fig. 3 Three-dimensional AI assurance by subarea, domain, and goal

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.

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.

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.

Appendix: All manuscripts and their detailed scores by ranking category

Year	Author	AI.s	R	M	Rs	Ds	Sz	Sc	L	G	A	C
1985	Ginsberg	KBS	1	1	1	1	1	1	0	0	1	0
1987	Castore	KBS	1	1	0	0	0	0	1	0	1	0
1987	Culbert	KBS	1	0	0	0	0	0	1	0	1	0
1987	O'Keefe	KBS	1	0	0	0	0	0	0	1	0	1
1988	Morell	KBS	1	1	1	0	0	1	1	0	0	0
1988	Sargent	ABS	1	1	0	0	0	0	0	1	0	1
1989	Becker	KBS	1	1	1	0	0	1	0	0	1	0
1990	Suen	KBS	1	1	1	0	0	0	0	0	0	0
1991	Vinze	KBS	1	1	1	0	0	1	0	0	0	0
1991	Gilstrap	KBS	1	1	0	0	0	0	0	1	0	0
1991	Raedt	KBS	1	1	0	0	0	0	0	0	0	1
1992	Andert	KBS	1	1	1	0	0	1	0	0	0	1
1992	Preece	KBS	1	1	1	0	0	0	0	0	0	1
1992	Davis	ABS	1	1	0	0	0	0	0	0	0	1
1993	Wells	KBS	1	1	0	0	0	1	0	1	0	0
1993	Mengshoel	KBS	1	1	0	0	0	1	0	0	0	0
1994	Zlatareva	KBS	1	1	1	0	0	1	0	0	1	0
1994	Lee	KBS	1	1	0	0	0	0	0	0	0	1
1995	Lockwood	KBS	1	1	1	0	0	1	0	0	1	0
1995	Singer	DS	1	1	1	1	0	1	0	0	0	0
1996	Martin	DS	0	1	1	0	0	1	0	1	1	1

Year	Author	AI.s	R	M	Rs	Ds	Sz	Sc	L	G	A	C
1996	Carley	KBS	1	0	0	0	0	0	0	1	0	1
1996	Gonzalez	KBS	1	1	0	0	0	1	0	0	0	0
1996	Abel	KBS	1	1	0	0	0	0	0	0	0	0
1996	Nourani	Generic	1	1	0	0	0	0	0	0	0	0
1997	Jones	GA	0	1	1	0	0	1	0	0	1	1
1997	Abel	KBS	1	1	1	0	0	0	0	0	0	0
1997	Harmelen	KBS	1	1	0	0	0	0	0	0	0	1
1997	Lowry	Generic	1	1	0	0	0	0	0	0	1	0
1997	Murrell	KBS	1	0	0	0	0	0	0	1	0	1
1997	Zaidi	KBS	1	1	0	0	0	1	0	0	0	0
1998	Miller	GA	1	1	1	0	0	1	0	0	0	0
1998	Zlatareva	KBS	1	1	1	0	0	1	0	0	0	0
1998	Antoniou	KBS	0	1	0	0	0	0	0	0	0	1
1999	Thomas	GA	1	1	1	1	1	1	0	0	0	1
1999	Tsai	KBS	1	1	0	0	0	0	0	0	1	1
1999	Hailu	RL	0	1	1	0	0	0	0	0	0	0
2000	Mosqueira-Rey	KBS	1	1	1	0	0	1	0	1	1	0
2000	Marcos	KBS	1	0	1	0	0	1	0	1	1	0
2000	Onoyama	KBS	1	1	1	0	0	1	0	0	1	0
2000	Edwards	DS	1	0	0	0	0	0	0	1	1	1
2000	Coenen	KBS	0	0	0	0	0	0	0	1	1	1
2000	Gonzalez	Generic	0	0	0	0	0	0	0	1	0	0
2001	Berndt	DS	1	1	1	1	1	1	0	0	1	0
2001	Halkidi	ML	1	1	1	0	0	1	0	0	1	1
2001	Barr	NLP	0	0	0	0	0	0	0	0	1	0
2002	Wu	KBS	1	1	0	0	0	1	1	0	1	1
2002	Knauf	KBS	1	1	0	0	0	1	0	0	0	0
2003	Schumann	DL	1	1	1	0	0	0	0	0	1	1
2003	Taylor	DL	1	1	0	0	0	0	0	1	0	1
2003	Kurd	DL	0	0	0	0	0	0	0	1	0	0
2004	Liu	ABS	1	0	0	0	0	0	0	1	0	1
2004	Sargent	ABS	1	0	0	0	0	0	0	1	0	1
2005	Liu	DL	1	1	1	1	0	1	0	0	0	1
2005	Menzies	Generic	1	1	0	0	0	0	0	1	0	0
2005	Taylor	DL	1	1	0	0	0	0	0	0	0	0
2006	Forster	AGI	1	1	1	1	1	1	0	0	0	1
2006	Taylor	DL	1	0	1	0	0	0	0	1	1	1
2006	Yilmaz	ABS	1	1	0	0	0	0	0	0	0	1
2006	Min	KBS	1	1	0	0	0	0	0	0	0	0
2006	Dibie-Barthélemy	KBS	0	0	0	0	0	0	0	0	0	1
2007	Martín-Guerrero	RL	1	1	1	1	0	1	0	0	1	0
2007	Knauf	KBS	1	1	1	1	0	0	0	0	0	1
2007	Brancovici	XAI	1	1	0	0	0	0	0	0	1	1
2007	Min	KBS	1	1	0	0	0	0	0	0	1	0
2008	Peppler	DS	1	1	1	1	1	1	1	0	1	0
2008	Liu	ABS	1	1	0	0	0	0	0	0	0	0
2009	Tadj	KBS	1	1	1	1	0	1	1	0	0	1
2009	Hibbard	AGI	0	1	1	0	0	1	0	0	0	0
2009	Pèpe	GA	0	1	1	0	0	0	0	0	0	0
2010	Bone	RL	1	1	1	1	1	1	0	0	1	0
2010	Winkler	CV	1	1	1	0	0	1	0	0	1	1

Year	Author	AI.s	R	M	Rs	Ds	Sz	Sc	L	G	A	C
2010	Dong	GA	1	1	1	0	0	1	0	0	0	1
2010	Niazi	ABS	1	1	1	1	0	1	0	0	0	0
2010	Sargent	ABS	0	0	0	0	0	0	0	1	0	1
2010	Schumann	DL	0	0	0	0	0	0	0	0	0	1
2012	Cohen	NLP	0	1	1	1	0	1	0	0	0	0
2012	Martinez-Balleste	CV	1	0	0	0	0	0	0	0	1	1
2013	Batarseh	KBS	1	1	1	1	1	1	0	0	1	0
2013	Sargent	ABS	1	1	1	0	0	0	0	0	1	1
2013	David	ABS	1	1	0	0	0	0	0	1	0	1
2013	Pitchforth	DL	1	1	0	0	0	0	0	0	0	0
2014	Ali	DS	1	1	1	1	1	1	1	0	0	1
2015	He	GA	1	1	1	0	0	1	1	0	1	1
2015	Rovcanin	RL	1	1	1	1	0	1	0	0	1	1
2015	Goodfellow	ML	1	1	1	0	0	1	0	0	0	1
2015	Arifin	ABS	1	0	0	0	0	0	0	1	1	1
2015	Batarseh	KBS	1	0	0	0	0	0	0	1	1	1
2015	Dobson	ML	1	0	0	0	0	0	0	0	0	1
2016	Veeramachaneni	DS	1	1	1	1	1	1	1	1	1	0
2016	Gao	DS	1	0	1	1	1	1	0	1	1	1
2016	Gulshan	CV	1	1	1	1	1	1	0	0	1	1
2016	Heaney	GA	1	1	1	1	1	1	1	0	1	0
2016	Aitken	ABS	1	1	1	1	1	1	0	0	0	1
2016	Celis	ML	1	1	1	1	1	1	0	0	1	0
2016	Jiang	RL	0	1	1	1	1	1	0	0	1	1
2016	Kianifar	GA	1	1	1	1	1	1	0	0	1	0
2016	Jilk	ABS	1	0	1	0	0	1	0	1	0	0
2016	Amodei	ML	1	0	0	0	0	0	0	1	0	1
2016	Aitken	ABS	1	1	0	0	0	0	0	0	0	0
2016	Janssen	DS	0	0	0	0	0	0	0	1	0	1
2016	Zeigler	ABS	1	1	0	0	0	0	0	0	0	0
2016	Wickramage	DS	1	0	0	0	0	0	0	0	0	0
2017	Liang	DS	1	1	1	1	0	1	0	0	1	1
2017	Xu	DL	1	1	1	1	1	1	0	0	0	1
2017	Mason	RL	1	1	1	0	0	1	0	0	0	1
2017	Leibovici	DS	1	1	0	0	0	0	0	1	0	1
2017	Laat	ML	0	1	0	0	0	0	0	1	0	1
2017	Mason	RL	1	1	0	0	0	0	0	0	0	1
2017	Lepri	ML	0	0	0	0	0	0	0	0	0	1
2018	Wei	RL	1	1	1	1	1	1	1	0	1	1
2018	Alves	ABS	1	1	1	1	1	1	0	0	1	1
2018	Elsayed	CV	1	1	1	1	1	1	0	0	0	1
2018	Go	DL	1	1	1	1	1	1	0	0	1	0
2018	Mason	RL	1	1	1	0	0	1	1	1	0	1
2018	Murray	XAI	1	1	1	1	1	1	0	0	0	1
2018	Pynadath	ML	1	1	1	0	0	1	1	0	1	1
2018	Stock	CV	1	1	1	1	1	1	0	0	1	0
2018	Cao	ML	1	1	1	1	0	1	0	0	0	1
2018	Ruan	DL	1	0	1	1	0	1	1	0	0	1
2018	Antunes	ML	1	1	1	0	0	1	1	0	0	0
2018	Volz	XAI	0	1	1	0	0	1	0	0	1	1
2018	AI Now	XAI	1	1	0	0	0	0	0	1	1	0

Year	Author	AI.s	R	M	Rs	Ds	Sz	Sc	L	G	A	C
2018	Došilović	ML	1	0	0	0	0	0	0	1	1	1
2018	EY	ML	1	0	0	0	0	0	0	1	1	1
2018	Guidotti	XAI	1	0	0	0	0	0	0	1	1	1
2018	Zhu	XAI	1	0	0	0	0	0	0	1	1	1
2018	Abdollahi	ML	1	0	0	0	0	0	0	1	0	1
2018	Adadi	XAI	1	0	0	0	0	0	0	1	0	1
2018	Agarwal	Generic	1	1	0	0	0	0	0	0	0	1
2018	Everitt	AGI	1	0	0	0	0	0	0	1	0	1
2018	Bride	XAI	1	1	0	0	0	0	0	0	0	0
2018	Hagras	XAI	1	0	0	0	0	0	0	0	0	1
2018	Kaul	ML	1	0	0	0	0	0	0	0	0	0
2018	Mehri	DL	0	0	0	0	0	0	0	0	0	1
2018	Sileno	XAI	1	0	0	0	0	0	0	0	0	0
2019	Tao	Generic	1	1	1	1	1	1	1	1	1	1
2019	Kaur	XAI	1	1	1	1	1	1	1	0	1	1
2019	Batarseh	DS	1	1	1	1	1	1	0	0	1	1
2019	Huber	RL	1	1	1	1	1	1	1	0	0	1
2019	Keneni	XAI	1	1	1	1	1	1	1	0	1	0
2019	Maloca	DL	1	1	1	1	1	1	0	0	1	1
2019	Barredo-Arrieta	XAI	1	1	1	1	1	0	0	1	1	0
2019	Chittajallu	XAI	1	1	1	1	0	1	1	0	0	1
2019	Ferreyra	XAI	1	1	1	0	0	1	0	1	1	1
2019	Lee	XAI	1	1	1	1	1	1	0	0	0	1
2019	Naqa	ML	1	1	1	1	0	1	0	0	1	1
2019	Prentzas	XAI	1	1	1	1	1	1	0	0	1	0
2019	Bellamy	XAI	1	1	1	0	0	1	1	1	0	0
2019	Beyret	RL	1	1	1	0	0	1	1	0	1	0
2019	Madumal	RL	1	1	1	0	0	1	0	0	1	1
2019	Schlegel	XAI	1	1	1	1	1	1	0	0	0	0
2019	Vabalas	ML	1	1	1	1	0	1	0	0	1	0
2019	Zhu	RL	1	1	1	0	0	1	1	0	0	1
2019	Chen	RL	1	1	0	0	0	0	0	1	1	1
2019	Cruz	RL	1	1	1	0	0	1	0	0	0	1
2019	Dupuis	XAI	1	1	1	0	0	1	0	0	0	1
2019	Joo	RL	1	1	1	0	0	1	1	0	0	0
2019	Ren	DL	0	1	1	0	0	1	1	0	1	0
2019	Srivastava	NLP	1	1	1	0	0	1	0	0	0	1
2019	Uslu	XAI	1	1	1	0	0	1	0	0	0	1
2019	Yoon	XAI	1	1	1	0	0	1	0	0	0	1
2019	Zhou	ML	1	1	1	0	0	1	0	0	0	1
2019	Mehrabi	ML	1	0	0	0	0	0	0	1	1	1
2019	Meskauskas	XAI	1	1	1	0	0	1	0	0	0	0
2019	Nassar	XAI	1	1	0	0	0	0	0	0	1	1
2019	Qiu	Generic	1	0	0	0	0	0	0	1	1	1
2019	Wang	XAI	1	1	0	0	0	0	0	1	0	1
2019	Breck	ML	1	1	0	0	0	0	0	0	1	0
2019	Glomsrud	XAI	1	0	0	0	0	0	0	0	1	1
2019	He	DL	1	0	0	0	0	0	0	1	0	1
2019	Israelsen	Generic	1	0	0	0	0	0	0	1	0	1
2019	Jha	DL	1	0	1	0	0	0	0	0	0	1
2019	Sun	XAI	0	1	1	0	0	1	0	0	0	0

Year	Author	AI.s	R	M	Rs	Ds	Sz	Sc	L	G	A	C
2019	Dghaym	XAI	0	0	0	0	0	0	0	0	1	1
2019	Mueller	XAI	1	0	0	0	0	0	0	0	0	1
2019	Protiviti	ML	0	0	0	0	0	0	0	1	0	1
2019	Spada	XAI	0	1	0	0	0	0	0	0	1	0
2019	Pocius	RL	1	0	0	0	0	0	0	0	0	0
2019	Rossi	XAI	1	0	0	0	0	0	0	0	0	0
2019	Varshney	ML	0	0	0	0	0	0	0	1	0	0
2020	D'Alterio	XAI	1	1	1	1	1	1	1	1	1	1
2020	Anderson	RL	1	1	1	1	1	1	0	1	1	1
2020	Birkenbihl	ML	1	1	1	1	1	1	1	0	1	1
2020	Checco	DS	1	1	1	1	1	1	0	1	1	1
2020	Chen	XAI	1	1	1	1	1	1	1	0	1	1
2020	EASA	DL	1	1	1	1	1	1	0	1	1	1
2020	Kulkarni	DS	1	1	1	1	1	1	0	1	1	1
2020	Kuppa	XAI	1	1	1	1	1	1	1	0	1	1
2020	Kuzlu	XAI	1	1	1	1	1	1	0	1	1	1
2020	Spinner	XAI	1	1	1	1	1	1	0	1	1	1
2020	Winkel	RL	1	1	1	1	1	1	1	0	1	1
2020	Gardiner	ML	1	1	1	1	1	1	0	0	1	1
2020	Guo	XAI	1	1	1	1	1	1	1	0	1	0
2020	Han	XAI	1	1	1	1	1	1	0	0	1	1
2020	Kohlbrenner	XAI	1	1	1	1	1	1	1	0	0	1
2020	Malolan	XAI	1	1	1	1	1	1	1	0	1	0
2020	Payrovnaziri	ML	1	1	0	1	1	0	1	1	1	1
2020	Sequeira	RL	1	1	1	1	1	1	0	0	1	1
2020	Sivamani	DL	1	1	1	1	1	1	0	0	1	1
2020	Tan	XAI	1	1	1	1	1	1	0	0	1	1
2020	Tao	XAI	1	1	1	1	1	1	1	0	1	0
2020	Welch	DL	1	1	1	1	1	1	0	0	1	1
2020	Xiao	DL	1	1	1	1	1	1	0	0	1	1
2020	Halliwell	DL	1	1	1	1	1	1	0	0	0	1
2020	Heuer	ML	1	1	1	1	1	1	0	0	0	1
2020	Kaur	XAI	1	1	1	1	1	1	0	0	1	0
2020	Mackowiak	CV	1	1	1	1	0	1	0	0	1	1
2020	Ragot	ML	1	1	1	1	1	1	0	0	1	0
2020	Rotman	RL	1	1	1	1	1	0	0	0	1	1
2020	Sarathy	XAI	1	1	1	1	0	1	0	0	1	1
2020	Uslu	XAI	0	1	1	1	1	1	0	0	1	1
2020	Cruz	RL	1	1	1	0	0	1	1	0	0	1
2020	He	RL	1	1	1	1	1	0	0	0	1	0
2020	Islam	XAI	0	1	1	1	1	0	0	0	1	1
2020	Mynuddin	RL	1	1	1	1	0	1	0	0	1	0
2020	Puiutta	RL	1	1	0	0	0	0	1	1	1	1
2020	Toreini	ML	1	0	0	0	1	1	1	1	0	1
2020	Toreini	ML	1	0	0	0	1	1	1	1	0	1
2020	Diallo	XAI	1	1	1	1	0	1	0	0	0	0
2020	Guo	XAI	1	0	0	0	0	1	0	1	1	1
2020	Haverinen	XAI	0	1	1	1	0	1	0	0	0	1
2020	Katell	XAI	1	1	0	0	0	0	1	0	1	1
2020	Murray	XAI	0	1	1	1	1	0	0	0	1	0
2020	Taylor	XAI	1	1	1	1	1	0	0	0	0	0

Year	Author	AI.s	R	M	Rs	Ds	Sz	Sc	L	G	A	C
2020	Tjoa	ML	1	0	0	0	1	0	1	1	0	1
2020	Varshney	ML	1	1	0	0	0	0	1	1	0	1
2020	Wieringa	XAI	1	0	0	0	1	0	1	1	0	1
2020	Wing	ML	1	0	0	0	1	0	1	1	0	1
2020	Das	XAI	1	0	0	0	0	0	1	1	0	1
2020	Li	XAI	0	0	0	0	1	0	1	1	0	1
2020	Dağlarli	XAI	1	0	0	0	0	0	0	1	0	1
2020	Dodge	XAI	1	1	0	0	0	0	0	0	0	1
2020	Heuillet	RL	1	0	0	0	0	0	0	1	0	1
2020	Martinez-Fernandez	XAI	1	1	0	0	0	0	0	0	1	0
2020	Putzer	XAI	0	1	0	0	0	0	0	0	1	1
2020	Raji	XAI	1	0	0	0	0	0	0	1	1	0
2020	Arrieta	XAI	1	0	0	0	0	0	0	0	0	1
2020	He	XAI	1	0	0	0	0	0	0	1	0	0
2020	Kaur	XAI	1	0	0	0	0	0	0	0	0	1
2020	Pawar	XAI	0	1	0	0	0	0	0	0	0	1
2020	Brennen	XAI	0	0	0	0	0	0	0	1	0	0
2020	European Commission	XAI	0	0	0	0	0	0	0	1	0	0
2021	Massoli	DL	1	1	1	1	1	1	1	0	1	1

Columns: AI subarea: AI.s; Relevance: R; Method: M; Results: Rs; Dataset: Ds; Size: Sz; Success: Sc; Limitations: L; General: G; Application: A; Comparison: C

Abbreviations
AI: Artificial Intelligence; TIMSS: Third International Math and Science Study; DS: Data Science; GA: Genetic Algorithms; RL: Reinforcement Learning; KBS: Knowledge-Based Systems; CV: Computer Vision; NLP: Natural Language Processing; ABS: Agent-Based Systems; ML: Machine Learning; DL: Deep Learning; XAI: Explainable AI.

Acknowledgements
Not applicable.

Authors' contributions
FB designed the study, provided the new assurance definition, developed the visualizations, and led the effort in writing the paper; LF reviewed the paper and provided consultation on the topic; CH developed the tables and the scoring system, and worked on finding, arranging, and managing the papers used in the review. All authors read and approved the final manuscript.

Funding
This work was supported by the Commonwealth Cyber Initiative (CCI).

Availability of data and materials
All data and materials are available under the following link: https://github.com/ferasbatarseh/AI-Assurance-Review.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]The Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University (Virginia Tech), Arlington, VA 22203, USA. [2]Department of Statistics, Virginia Polytechnic Institute and State University (Virginia Tech), Arlington, VA 22203, USA. [3]College of Science, George Mason University, Fairfax 22030, USA.

Received: 13 January 2021 Accepted: 22 March 2021
Published online: 26 April 2021

References

1. Abdollahi B, Nasraoui O. Transparency in fair machine learning: the case of explainable recommender systems. In: Zhou J, Chen F, editors. Human and machine learning: visible, explainable, trustworthy and transparent. Berlin: Springer; 2018 https://doi.org/10.1007/978-3-319-90403-0_2.
2. Abel T, Gonzalez A (1997). Utilizing Criteria to Reduce a Set of Test Cases for Expert System Validation.
3. Abel T, Knauf R, Gonzalez A. (1996). Generation of a minimal set of test cases that is functionally equivalent to an exhaustive set, for use in knowledge-based system validation.
4. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access. 2018;6:23.
5. Agarwal A., Lohia P, Nagar, S, Dey K, Saha D. (2018). Automated Test Generation to Detect Individual Discrimination in AI Models ArXiv:1809.03260 [Cs].
6. Aitken M. Assured human-autonomy interaction through machine self-confidence. Colorado: University of Colorado; 2016.
7. Algorithmic Accountability Policy Tooklit. (2018). AI NOW.
8. Ali AL, Schmid F. Data quality assurance for volunteered geographic information. In: Duckham M, Pebesma E, Stewart K, Frank AU, editors. Geographic information science. Berlin: Springer; 2014. p. 126–41.
9. Alves E, Bhatt D, Hall B, Driscoll K, Murugesan A (2018). Considerations in Assuring Safety of Increasingly Autonomous Systems (NASA Contractor Report NASA/CR–2018–22008; Issue NASA/CR–2018–22008). NASA.
10. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. Concrete Problems in AI Safety. ArXiv:1606.06565[Cs].; 2016.
11. Anderson A, Dodge J, Sadarangani A, Juozapaitis Z, Newman E, Irvine J, Chattopadhyay S, Olson M, Fern A, Burnett M. Mental models of mere mortals with explanations of reinforcement learning. ACM Trans Interact Intell Syst. 2020;10(2):1–37. https://doi.org/10.1145/3366485.
12. Andert EP. Integrated knowledge-based system design and validation for solving problems in uncertain environments. Int J Man Mach Stud. 1992;36(2):357–73. https://doi.org/10.1016/0020-7373(92)90023-E.
13. Antoniou G, Harmelen F, Plant R, Vanthienen J. Verification and validation of knowledge-based systems: report on two 1997 events. AI Mag. 1998;19:123–6.
14. Antunes N, Balby L, Figueiredo F, Lourenco N, Meira W, Santos W (2018). Fairness and Transparency of Machine Learning for Trustworthy Cloud Services. 2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W), 188–193. https://doi.org/10.1109/DSN-W.2018.00063
15. Arifin SMN, Madey GR. Verification, validation, and replication methods for agent-based modeling and simulation: lessons learned the hard way! In: Yilmaz L, editor. Concepts and methodologies for modeling and simulation: a tribute to Tuncer Ören. Berlin: Springer; 2015. p. 217–42.
16. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R, Chatila R, Herrera F. (2019). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI.ArXiv:1910.10045 [Cs].
17. Assurance in the age of AI. (2018). EY.
18. Barr VB, Klavans JL. Verification and validation of language processing systems: is it evaluation? Proc Workshop Eval Lang Dialogue Syst. 2001;9:1–7. https://doi.org/10.3115/1118053.1118058.
19. Barredo-Arrieta A, Lana I, Del Ser J. What lies beneath: a note on the explainability of black-box machine learning models for road traffic forecasting. IEEE Intell Transp Syst Conf (ITSC). 2019;2019:2232–7. https://doi.org/10.1109/ITSC.2019.8916985.
20. Batarseh FA, Gonzalez AJ. Incremental lifecycle validation of knowledge-based systems through commonKADS. EEE Trans Syst Man Cybern. 2013;43(3):12.
21. Batarseh FA, Gonzalez AJ. Validation of knowledge-based systems: a reassessment of the field. Artif Intell Rev. 2015;43(4):485–500. https://doi.org/10.1007/s10462-013-9396-9.
22. Batarseh AF, Yang R. Transforming Government and Agricultural Policy Using Artificial Intelligence: Federal Data Science; 2017.
23. Batarseh A, Feras, Mohod R, Kumar, A, and Bui J. Chapter 10: the Application of Artificial Intelligence in Software Engineering: a Review Challenging Conventional Wisdom. (2020). In Data Democracy, Elsevier Academic Press. pp. 179–232
24. Batarseh F. A, Kulkarni A. (2019). Context-Driven Data Mining through Bias Removal and Incompleteness Mitigation. 7.
25. Becker L. A, Green P. G, Bhatnagar J. (1989). Evidence Flow Graph Methods for Validation and Verification of Expert Systems (NASA Contractor Report No. 181810; p. 46). Worcester Polytechnic Institute.
26. Bellamy RKE, Mojsilovic A, Nagar S, Ramamurthy KN, Richards J, Saha D, Sattigeri P, Singh M, Varshney KR, Zhang Y, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, Lohia P, Martino J, Mehta S. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM J Res Dev. 2019;63(4):1–4. https://doi.org/10.1147/JRD.2019.2942287.
27. Berndt DJ, Fisher JW, Hevner AR, Studnicki J. Healthcare data warehousing and quality assurance. Computer. 2001;34(12):56–65. https://doi.org/10.1109/2.970578.
28. Beyret B, Shafti A, Faisal AA. Dot-to-Dot: explainable hierarchical reinforcement learning for robotic manipulation. IEEE/RSJ Int Conf Intell Robots Syst (IROS). 2019;2019:5014–9. https://doi.org/10.1109/IROS40897.2019.8968488.
29. Birkenbihl C. Differences in cohort study data affect external validation of artificial intelligence models for predictive diagnostics of dementia—Lessons for translation into clinical practice. EPMA J. 2020;11(3):367–76.
30. Bone C, Dragićević S. Simulation and validation of a reinforcement learning agent-based model for multi-stakeholder forest management. Comput Environ Urban Syst. 2010;34(2):162–74. https://doi.org/10.1016/j.compenvurbsys.2009.10.001.
31. Brancovici, G. (2007). Towards Trustworthy Intelligence on the Road: A Flexible Architecture for Safe, Adaptive, Autonomous Applications. 2007 IEEE Congress on Evolutionary Computation, Singapore. https://doi.org/10.1109/CEC.2007.4425023

32. Breck E, Zinkevich M, Polyzotis N, Whang S, Roy S. (2019). Data Validation for Machine Learning. Proceedings of SysML. https://mlsys.org/Conferences/2019/doc/2019/167.pdf
33. Brennen, A. (2020). What Do People Really Want When They Say They Want "Explainable AI?"We Asked 60 Stakeholders. Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, 1–7. https://doi.org/10.1145/3334480.3383047
34. Bride H, Dong J. S, Hóu Z, Mahony B, Oxenham M. (2018). Towards Trustworthy AI for Autonomous Systems. In J. Sun M. Sun (Eds.), Formal Methods and Software Engineering (pp. 407–411). Springer International Publishing. https://doi.org/10.1007/978-3-030-02450-5_24
35. Cao N, Li G, Zhu P, Sun Q, Wang Y, Li J, Yan M, Zhao Y. Handling the adversarial attacks. J Ambient Intell Humaniz Comput. 2019;10(8):2929–43. https://doi.org/10.1007/s12652-018-0714-6.
36. Carley K. M. (1996). Validating Computational Models [Work Paper]. Carnegie Mellon University.
37. Castore G. (1987). A Formal Approach to Validation and Verification for Knowledge-Based Control. Systems. 6.
38. Celis L. E, Deshpande A, Kathuria T, Vishnoi N. K. (2016). How to be Fair and Diverse? ArXiv:1610.07183 [Cs].
39. Checco A, Bates J, Demartini G. Adversarial attacks on crowdsourcing quality control. J Artif Intell Res. 2020;67:375–408. https://doi.org/10.1613/jair.1.11332.
40. Chen H-Y, Lee C-H. Vibration signals analysis by explainable artificial intelligence (XAI) approach: application on bearing faults diagnosis. IEEE Access. 2020;8:134246–56. https://doi.org/10.1109/ACCESS.2020.3006491.
41. Chen T, Liu J, Xiang Y, Niu W, Tong E, Han Z. Adversarial attack and defense in reinforcement learning-from AI security view. Cybersecurity. 2019;2(1):11. https://doi.org/10.1186/s42400-019-0027-x.
42. Chittajallu, D. R, Dong B, Tunison P, Collins R, Wells K, Fleshman J, Sankaranarayanan G, Schwaitzberg S, Cavuoto L, Enquobahrie A. (2019). XAI-CBIR: Explainable AI System for Content based Retrieval of Video Frames from Minimally Invasive Surgery Videos. 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), 66–69. https://doi.org/10.1109/ISBI.2019.8759428
43. Cluzeau J. M, Henriquel X, Rebender G, Soudain G, Dijk L. van, Gronskiy A, Haber D, Perret-Gentil C, Polak R. (2020). Concepts of Design Assurance for Neural Networks (CoDANN) [Public Report Extract]. European Union Aviation Safety Agency.
44. Coenen F, Bench-Capon T, Boswell R, Dibie-Barthélemy J, Eaglestone B, Gerrits R, Grégoire E, Lige¸za, A, Laita, L, Owoc, M, Sellini, F, Spreeuwenberg, S, Vanthienen, J, Vermesan, A, Wiratunga, N. . Validation and verification of knowledge-based systems: report on EUROVAV99. Knowl Eng Rev. 2000;15(2):187–96. https://doi.org/10.1017/S0269888900002010.
45. Cohen KB, Hunter LE, Palmer M. Assessment of software testing and quality assurance in natural language processing applications and a linguistically inspired approach to improving it. In: Moschitti A, Plank B, editors. Trustworthy eternal systems via evolving software, data and knowledge. Berlin: Springer; 2013. p. 77–90. https://doi.org/10.1007/978-3-642-45260-4_6
46. Cruz F, Dazeley R, Vamplew P. Memory-Based Explainable Reinforcement Learning. In: Liu J, Bailey J, editors. AI 2019: Advances in Artificial Intelligence, vol. 11919. Berlin: Springer; 2019. p. 66–77.https://doi.org/10.1007/978-3-030-35288-2_6
47. Cruz F, Dazeley R, Vamplew P. (2020). Explainable robotic systems: Understanding goal-driven actions in a reinforcement learning scenario. ArXiv:2006.13615[Cs].nn
48. Culbert, C, Riley G, Savely R. T. (1987). Approaches to the Verification of Rule-Based Expert Systems. SOAR'87L First Annual Workshop on Space Operation Automation and Robotics, 27–37.
49. Dağlarli E. (2020). Explainable Artificial Intelligence (xAI) Approaches and Deep Meta-Learning Models. In Advances and Applications in Deep Learning. IntechOpen.
50. D'Alterio, P Garibaldi, J. M John, R. I. (2020). Constrained Interval Type-2 Fuzzy Classification Systems for Explainable AI (XAI). 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 1–8. https://doi.org/10.1109/FUZZ48607.2020.9177671
51. Das A, Rad P. (2020). Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. ArXiv:2006.11371[Cs].n
52. David, N. (2013). Validating Simulations. In Simulating Social Complexity (pp. 135–171). Springer Berlin Heidelberg.
53. Davis P. K. (1992). Generalizing concepts and methods of verification, validation, and accreditation (VV&A) for military simulations. Rand.
54. de Laat PB. Algorithmic decision-making based on machine learning from big data: can transparency restore accountability? Philos Technol. 2018;31(4):525–41. https://doi.org/10.1007/s13347-017-0293-z.
55. De Raedt L, Sablon G, Bruynooghe M. Using Interactive Concept Learning for Knowledge-base Validation and Verification. In: Validation, verification and test of knowledge-based systems. Hoboken: Wiley; 1991. p. 177–90.
56. Dghaym D, Turnock S, Butler M, Downes J, Hoang T. S, Pritchard B. (2020). Developing a Framework for Trustworthy Autonomous Maritime Systems. In Proceedings of the International Seminar on Safety and Security of Autonomous Vessels (ISSAV) and European STAMP Workshop and Conference (ESWC) 2019 (pp. 73–82). Sciendo. https://doi.org/10.2478/9788395669606-007
57. Diallo, A. B, Nakagawa H, Tsuchiya T. (2020). An Explainable Deep Learning Approach for Adaptation Space Reduction. 2020 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C), 230–231. https://doi.org/10.1109/ACSOS-C51401.2020.00063
58. Dibie-Barthelemy J, Haemmerle O, Salvat E. (2006). A semantic validation of conceptual graphs. 13.
59. Dobson J. Can an algorithm be disturbed?: Machine learning, intrinsic criticism, and the digital humanities. Coll Lit. 2015;42:543–64. https://doi.org/10.1353/lit.2015.0037.
60. US Department of Defense (DoD) Directive 5000.59. 1995.
61. Dodge J, Burnett M. (2020). Position: We Can Measure XAI Explanations Better with Templates. ExSS-ATEC@IUI, 1–13.
62. Dong G, Wu S, Wang G, Guo T, Huang Y. Security assurance with metamorphic testing and genetic algorithm. IEEE/WIC/ACM Int Conf Web Intell Agent Technol. 2010;2010:397–401. https://doi.org/10.1109/WI-IAT.2010.101.

63. Došilović, F. K, Brcic M, Hlupic N. (2018). Explainable artificial intelligence: A survey. 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 0210–0215. https://doi.org/10.23919/MIPRO.2018.8400040
64. Dupuis NK, Verheij DB. An analysis of decompositional rule extraction for explainable neural Networks. Groningen: University of Groningen; 2019.
65. Edwards D. Data Quality Assurance. In: Ecological data: design, management and processing. Hoboken: Blackwell; 2000. p. 70–91.
66. El Naqa I, Irrer J, Ritter TA, DeMarco J, Al-Hallaq H, Booth J, Kim G, Alkhatib A, Popple R, Perez M, Farrey K, Moran JM. Machine learning for automated quality assurance in radiotherapy: a proof of principle using EPID data description. Med Phys. 2019;46(4):1914–21. https://doi.org/10.1002/mp.13433.
67. Elsayed G, Shankar S, Cheung B, Papernot N, Kurakin A, Goodfellow I, Sohl-Dickstein J. (2018). Adversarial Examples that Fool both Computer Vision and Time-Limited Humans. 11.
68. Everitt T, Lea G, Hutter M. (2018). AGI Safety Literature Review. ArXiv:1805.01109[Cs].
69. Ferreyra E, Hagras H, Kern M, Owusu G. (2019). Depicting Decision-Making: A Type-2 Fuzzy Logic Based Explainable Artificial Intelligence System for Goal-Driven Simulation in the Workforce Allocation Domain. 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 1–6. https://doi.org/10.1109/FUZZ-IEEE.2019.8858933
70. Forster D. A. (2006). Validation of individual consciousness in Strong Artificial Intelligence: An African Theological contribution. University of South Africa.
71. Gao J, Xie C, Tao C. Big data validation and quality assurance—issuses, challenges, and Needs. IEEE Symposium on Service-Oriented System Engineering (SOSE). 2016;2016:433–41. https://doi.org/10.1109/SOSE.2016.63.
72. Gardiner L-J, Carrieri AP, Wilshaw J, Checkley S, Pyzer-Knapp EO, Krishna R. Using human in vitro transcriptome analysis to build trustworthy machine learning models for prediction of animal drug toxicity. Sci Rep. 2020;10(1):9522. https://doi.org/10.1038/s41598-020-66481-0.
73. Gilstrap L. Validation and verification of expert systems. Telematics Inform. 1991;8(4):439–48. https://doi.org/10.1016/S0736-5853(05)80064-4.
74. Ginsberg A, Weiss S. (2001). SEEK2: A Generalized Approach to Automatic Knowledge Base Refinement. 9th International Joint Conference on Artificial Intelligence, 1, 8.
75. Glomsrud J. A, Ødegårdstuen A, Clair A. L. S, Smogeli Ø. (2020). Trustworthy versus Explainable AI in Autonomous Vessels. In Proceedings of the International Seminar on Safety and Security of Autonomous Vessels (ISSAV) and European STAMP Workshop and Conference (ESWC) 2019 (pp. 37–47). Sciendo. https://doi.org/10.2478/9788395669606-004
76. Go, W Lee D. (2018). Toward Trustworthy Deep Learning in Security. Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, 2219–2221. https://doi.org/10.1145/3243734.3278526
77. Gonzalez AJ, Barr V. Validation and verification of intelligent systems—What are they and how are they different? J Exp Theor Artif Intell. 2000;12(4):407–20. https://doi.org/10.1080/095281300454793.
78. Gonzalez AJ, Gupta UG, Chianese RB. Performance evaluation of a large diagnostic expert system using a heuristic test case generator. Eng Appl Artif Intell. 1996;9(3):275–84. https://doi.org/10.1016/0952-1976(95)00018-6.
79. Goodfellow I. J, Shlens J, Szegedy C. (2015). Explaining and Harnessing Adversarial Examples. ArXiv:1412.6572 [Cs, Stat].nn
80. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv. 2019;51(5):1–42. https://doi.org/10.1145/3236009.
81. Gulshan V, Peng L, Coram, M, Stumpe M. C, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson P. C, Mega J. L, Webster D. R. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. 9.
82. Guo W. Explainable artificial intelligence for 6G: improving trust between human and machine. IEEE Commun Mag. 2020;58(6):39–45. https://doi.org/10.1109/MCOM.001.2000050.
83. Hagras H. Toward human-understandable, explainable AI. Computer. 2018;51(9):28–36. https://doi.org/10.1109/MC.2018.3620965.
84. Hailu G, Sommer G. (1999). On amount and quality of bias in reinforcement learning. IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.99CH37028), 2, 728–733. https://doi.org/10.1109/ICSMC.1999.825352
85. Halkidi M, Batistakis Y, Vazirgiannis M. On clustering validation techniques. J Intell Inf Syst. 2001;17(2/3):107–45.
86. Halliwell N, Lecue F. (2020). Trustworthy Convolutional Neural Networks: A Gradient Penalized-based Approach. ArXiv:2009.14260[Cs].
87. Han S-H, Kwon M-S, Choi H-J. EXplainable AI (XAI) approach to image captioning. J Eng. 2020;2020(13):589–94. https://doi.org/10.1049/joe.2019.1217.
88. Harmelen F, Teije A. (1997). Validation and Verification of Conceptual Models of Diagnosis. Fourth European Symposium on the Validation and Verification of Knowledge-Based Systems, 117–128.
89. Haverinen T. (2020). Towards Explainable Artificial Intelligence (XAI) [Master's Thesis]. University of Jyväskylä.
90. He C, Xing J, Li J, Yang Q, Wang R, Zhang X. A new optimal sensor placement strategy based on modified modal assurance criterion and improved adaptive genetic algorithm for structural health monitoring. Math Probl Eng. 2015;2015:1–10. https://doi.org/10.1155/2015/626342.
91. He H, Gray J, Cangelosi A, Meng Q, McGinnity T. M, Mehnen J. (2020). The Challenges and Opportunities of Artificial Intelligence for Trustworthy Robots and Autonomous Systems. 2020 3rd International Conference on Intelligent Robotic and Control Engineering (IRCE), 68–74. https://doi.org/10.1109/IRCE50905.2020.9199244
92. He Y, Meng G, Chen K, Hu X, He J. (2020). Towards Security Threats of Deep Learning Systems: A Survey. ArXiv:1911.12562[Cs].
93. Heaney KD, Lermusiaux PFJ, Duda TF, Haley PJ. Validation of genetic algorithm-based optimal sampling for ocean data assimilation. Ocean Dyn. 2016;66(10):1209–29. https://doi.org/10.1007/s10236-016-0976-5.

94. Heuer H, Breiter A. (2020). More Than Accuracy: Towards Trustworthy Machine Learning Interfaces for Object Recognition. Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, 298–302. https://doi.org/10.1145/3340631.3394873
95. Heuillet A, Couthouis F, Díaz-Rodríguez N. (2020). Explainability in Deep Reinforcement Learning. ArXiv:2008.06693 [Cs].
96. Hibbard B. Bias and no free lunch in formal measures of intelligence. J Artif General Intell. 2009;1(1):54–61. https://doi.org/10.2478/v10229-011-0004-6.
97. Huber T. (2019). Enhancing Explainability of Deep Reinforcement Learning Through Selective Layer-Wise Relevance Propagation. 15.
98. Islam MA, Anderson DT, Pinar A, Havens TC, Scott G, Keller JM. Enabling explainable fusion in deep learning with fuzzy integral neural Networks. IEEE Trans Fuzzy Syst. 2019. https://doi.org/10.1109/TFUZZ.2019.2917124.
99. Israelsen B. W, Ahmed N. R. (2019). "Dave...I can assure you ...that it's going to be all right ..." A Definition, Case for, and Survey of Algorithmic Assurances in Human-Autonomy Trust Relationships. ACM Computing Surveys, 51(6), 1–37. https://doi.org/10.1145/3267338
100. Janssen M, Kuk G. The challenges and limits of big data algorithms in technocratic governance. Gov Inf Q. 2016;33(3):371–7. https://doi.org/10.1016/j.giq.2016.08.011.
101. Jha S, Raj S, Fernandes S, Jha S. K, Jha S, Jalaian B, Verma G, Swami A. (2019). Attribution-Based Confidence Metric For Deep Neural Networks. https://openreview.net/forum?id=rkeYFrHgIB
102. Jiang N, Li L. (2016). Doubly Robust Off-policy Value Evaluation for Reinforcement Learning. 33 Rd International Conference on Machine Learning, 48, 10.
103. Jilk, D. J. (2018). Limits to Verification and Validation of Agentic Behavior. In Artificial Intelligence Safety and Security (pp. 225–234). Taylor Francis Group. https://doi.org/10.1201/9781351251389-16
104. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking11 Edited by F E Cohen. J Mol Biol. 1997;267(3):727–48. https://doi.org/10.1006/jmbi.1996.0897.
105. Jorge E, Brynte L, Cronrath C, Wigstrom O, Bengtsson K, Gustavsson E, Lennartson B, Jirstrand M. Reinforcement learning in real-time geometry assurance. In: 51st CIRP Proceedings of the Conference on Manufacturing Systems. 2018. p. 1073–8.
106. Joo H-T, Kim K-J. Visualization of deep reinforcement learning using Grad-CAM: how AI plays atari games? IEEE Conf Games (CoG). 2019;2019:1–2. https://doi.org/10.1109/CIG.2019.8847950.
107. Katell M, Young M, Dailey D, Herman B, Guetler V, Tam A, Binz C, Raz D, Krafft P. M. (2020). Toward situated interventions for algorithmic equity: Lessons from the field. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 45–55. https://doi.org/10.1145/3351095.3372874
108. Kaul S. (2018). Speed And Accuracy Are Not Enough! Trustworthy Machine Learning. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 372–373. https://doi.org/10.1145/3278721.3278796
109. Kaur D, Uslu S, Durresi A. Trust-based security mechanism for detecting clusters of fake users in social networks. In: Barolli L, Takizawa M, Xhafa F, Enokido T, editors. Web, artificial intelligence and network applications, vol. 927. Berlin: Springer; 2019. p. 641–50. https://doi.org/10.1007/978-3-030-15035-8_62.
110. Kaur D, Uslu S, Durresi A. Requirements for Trustworthy Artificial Intelligence – A Review. In: Barolli L, Li KF, Enokido T, Takizawa M, editors. Advances in Networked-Based Information Systems, vol. 1264. Berlin: Springer; 2021. p. 105–15; https://doi.org/10.1007/978-3-030-57811-4_11.
111. Kaur D, Uslu S, Durresi A, Mohler G, Carter JG. Trust-based human-machine collaboration mechanism for predicting crimes. In: Barolli L, Amato F, Moscato F, Enokido T, Takizawa M, editors. Advanced information networking and applications, vol. 1151. Berlin: Springer; 2020. p. 603–16; https://doi.org/10.1007/978-3-030-44041-1_54.
112. Keneni BM, Kaur D, Al Bataineh A, Devabhaktuni VK, Javaid AY, Zaientz JD, Marinier RP. Evolving rule-based explainable artificial intelligence for unmanned aerial vehicles. IEEE Access. 2019;7:17001–16. https://doi.org/10.1109/ACCESS.2019.2893141.
113. Kianifar MR. Application of permutation genetic algorithm for sequential model building–model validation design of experiments. Soft Comput. 2016;20:3023–44. https://doi.org/10.1007/s00500-015-1929-5.
114. Knauf R, Gonzalez AJ, Abel T. A framework for validation of rule-based systems. Cybern PART B. 2002;32(3):15.
115. Knauf R, Tsuruta S, Gonzalez AJ. Toward reducing human involvement in validation of knowledge-based systems. IEEE Trans Syst Man Cybern Part A. 2007;37(1):120–31. https://doi.org/10.1109/TSMCA.2006.886365.
116. Kohlbrenner M, Bauer A, Nakajima S, Binder A, Samek W, Lapuschkin S. Towards best practice in explaining neural network decisions with LRP. Int Joint Conf Neural Netw. 2020;2020:1–7. https://doi.org/10.1109/IJCNN48605.2020.9206975.
117. Kulkarni A, Chong D, Batarseh FA. Foundations of data imbalance and solutions for a data democracy. In: data democracy. Cambridge: Academic Press; 2020. p. 83–106.
118. Kuppa A, Le-Khac N-A. Black box attacks on explainable artificial intelligence (XAI) methods in cyber security. Int Joint Confer Neural Netw. 2020;2020:1–8. https://doi.org/10.1109/IJCNN48605.2020.9206780.
119. Kurd Z, Kelly T. Safety lifecycle for developing safety critical artificial neural networks. In: Anderson S, Felici M, Littlewood B, editors. Computer safety, reliability, and security. Berlin: Springer; 2003. p. 77–91.
120. Kuzlu M, Cali U, Sharma V, Guler O. Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools. IEEE Access. 2020;8:187814–23. https://doi.org/10.1109/ACCESS.2020.3031477.
121. Lee J, ha, Shin, I. hee, Jeong, S. gu, Lee, S.-I, Zaheer, M. Z, Seo, B.-S. . Improvement in deep networks for optimization using eXplainable artificial intelligence. 2019 2019 International Conference on Information and Communication Technology Convergence (ICTC), 525–30. https://doi.org/10.1109/ICTC46691.2019.8939943.
122. Lee S, O'Keefe RM. Developing a strategy for expert system verification and validation. IEEE Trans Syst Man Cybern. 1994;24(4):643–55. https://doi.org/10.1109/21.286384.
123. Leibovici D. G, Rosser J. F, Hodges C, Evans B, Jackson M. J, Higgins C. I. (2017). On Data Quality Assurance and Conflation Entanglement in Crowdsourcing for Environmental Studies. 17.

124. Lepri B, Oliver N, Letouzé E, Pentland A, Vinck P. Fair, transparent, and accountable algorithmic decision-making processes: the premise, the proposed solutions, and the open challenges. Philos Technol. 2018;31(4):611–27. https://doi.org/10.1007/s13347-017-0279-x.
125. Li X-H, Cao CC, Shi Y, Bai W, Gao H, Qiu L, Wang C, Gao Y, Zhang S, Xue X, Chen L. A survey of data-driven and knowledge-aware eXplainable AI. IEEE Trans Knowl Data Eng. 2020. https://doi.org/10.1109/TKDE.2020.2983930.
126. Liang X, Zhao J, Shetty S, Li D. (2017). Towards data assurance and resilience in IoT using blockchain. MILCOM 2017 - 2017 IEEE Military Communications Conference (MILCOM), 261–266. https://doi.org/10.1109/MILCOM.2017.8170858
127. Liu F, Yang M. (2004). Verification and validation of al simulation systems. Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.04EX826), 3100–3105. https://doi.org/10.1109/ICMLC.2004.1378566
128. Liu F, Yang M. (2005). Verification and Validation of Artificial Neural Network Models AI 2005: Advances in Artificial Intelligence, 3809:1041–1046.
129. Liu F, Yang M, Shi P. (2008). Verification and validation of fuzzy rules-based human behavior models. 2008 Asia Simulation Conference - 7th International Conference on System Simulation and Scientific Computing, 813–819. https://doi.org/10.1109/ASC-ICSC.2008.4675474
130. Lockwood S, Chen Z. Knowledge validation of engineering expert systems. Adv Eng Softw. 1995;23(2):97–104. https://doi.org/10.1016/0965-9978(95)00018-R.
131. Lowry M, Havelund K, Penix J. Verification and validation of AI systems that control deep-space spacecraft. In: Raś ZW, Skowron A, editors. Foundations of Intelligent Systems, vol. 1325. Berlin: Springer; 1997. p. 35–47; https://doi.org/10.1007/3-540-63614-5_3.
132. Mackowiak R, Ardizzone L, Köthe U, Rother, C. (2020). Generative Classifiers as a Basis for Trustworthy Computer Vision. ArXiv:2007.15036 [Cs].nn
133. Madumal P, Miller T, Sonenberg L, Vetere F. (2019). Explainable Reinforcement Learning Through a Causal Lens. ArXiv:1905.10958 [Cs, Stat].
134. Maloca PM, Lee AY, de Carvalho ER, Okada M, Fasler K, Leung I, Hörmann B, Kaiser P, Suter S, Hasler PW, Zarranz-Ventura J, Egan C, Heeren TFC, Balaskas K, Tufail A, Scholl HPN. Validation of automated artificial intelligence segmentation of optical coherence tomography images. PLoS ONE. 2019;14(8):e0220063. https://doi.org/10.1371/journal.pone.0220063.
135. Magazzeni D, McBurney P, Nash W. Validation and Verification of Smart Contracts: A Research Agenda. Computer. 2017;50(9):50–57. https://doi.org/10.1109/MC.2017.3571045
136. Malolan B, Parekh A, Kazi F. (2020). Explainable Deep-Fake Detection Using Visual Interpretability Methods. 2020 3rd International Conference on Information and Computer Technologies (ICICT), 289–293. https://doi.org/10.1109/ICICT50521.2020.00051
137. Marcos M, del Pobil AP, Moisan S. Model-based verification of knowledge-based systems: a case study. IEE Proceedings - Software. 2000;147(5):163. https://doi.org/10.1049/ip-sen:20000896.
138. Martin M. O, Mullis I. V. S, Bruneforth M, Third International Mathematics and Science Study (Eds.). (1996). Quality assurance in data collection. Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.
139. Martinez-Balleste, A, Rashwan, H. A, Puig, D, Fullana, A. P. (2012). Towards a trustworthy privacy in pervasive video surveillance systems. 2012 IEEE International Conference on Pervasive Computing and Communications Workshops, 914–919. https://doi.org/10.1109/PerComW.2012.6197644
140. Martínez-Fernández S, Franch X, Jedlitschka A, Oriol M, Trendowicz A. (2020). Research Directions for Developing and Operating Artificial Intelligence Models in Trustworthy Autonomous Systems. ArXiv:2003.05434[Cs].n
141. Martín-Guerrero JD, Soria-Olivas E, Martínez-Sober M, Climente-Martí M, De Diego-Santos T, Jiménez-Torres NV. Validation of a reinforcement learning policy for dosage optimization of erythropoietin. In: Orgun MA, Thornton J, editors. AI 2007: Advances in artificial intelligence. Berlin: Springer; 2007. p. 732–8.
142. Mason G, Calinescu R, Kudenko D, Banks A. (2017a). Assured Reinforcement Learning for Safety-Critical Applications.
143. Mason G, Calinescu R, Kudenko D, Banks A. Assurance in reinforcement learning using quantitative verification. In: Hatzilygeroudis I, Palade V, editors. Advances in hybridization of intelligent methods, vol. 85. Berlin: Springer; 2018. p. 71–96.
144. Mason G, Calinescu R, Kudenko D, Banks A. (2017b). Assured Reinforcement Learning with Formally Verified Abstract Policies. Proceedings of the 9th International Conference on Agents and Artificial Intelligence, 105–117. https://doi.org/10.5220/0006156001050117
145. Massoli FV, Carrara F, Amato G, Falchi F. Detection of face recognition adversarial attacks. Comput Vision Image Understand. 2021;11:103103.
146. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. (2019). A Survey on Bias and Fairness in Machine Learning. ArXiv:1908.09635 [Cs].n
147. Mehri V. A, Ilie D, Tutschku K. (2018). Privacy and DRM Requirements for Collaborative Development of AI Applications. Proceedings of the 13th International Conference on Availability, Reliability and Security - ARES 2018, 1–8. https://doi.org/10.1145/3230833.3233268
148. Mengshoel OJ. Knowledge validation: principles and practice. IEEE Expert. 1993;8(3):62–8. https://doi.org/10.1109/64.215224.
149. Menzies T, Pecheur C. Verification and validation and artificial intelligence. In: Advances in computers, vol. 65. Amsterdam: Elsevier; 2005. p. 153–201.
150. Meskauskas Z, Jasinevicius R, Kazanavicius E, Petrauskas, V. (2020). XAI-Based Fuzzy SWOT Maps for Analysis of Complex Systems. 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 1–8. https://doi.org/10.1109/FUZZ48607.2020.9177792
151. Miller J. Active Nonlinear Test (ANTs) of Complex Simulation Models. Manag Sci. 1998;44(6):482.
152. Min F, Ma P, Yang M. A knowledge-based method for the validation of military simulation. Winter Simulation Conf. 2007;2007:1395–402. https://doi.org/10.1109/WSC.2007.4419748.

153. Min Fei-yan, Yang, M, Wang, Z. (2006). An Intelligent Validation System of Simulation Model. 2006 International Conference on Machine Learning and Cybernetics, 1459–1464. https://doi.org/10.1109/ICMLC.2006.258759
154. Morell L. J. (1988). Use of metaknowledge in the verification of knowledge-based systems. Proceedings of the 1st International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems - Volume 2, 847–857. https://doi.org/10.1145/55674.55699
155. Mosqueira-Rey E, Moret-Bonillo V. Validation of intelligent systems: a critical study and a tool. Expert Syst Appl. 2000;16:1–6.
156. Mueller ST, Hoffman, RR, Clancey W, Emrey A, Klein G. (2019). Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI. ArXiv:1902.01876 [Cs].n
157. Murray B, Islam M. A, Pinar A. J, Havens, T. C, Anderson D. T, Scott G. (2018). Explainable AI for Understanding Decisions and Data-Driven Optimization of the Choquet Integral. 2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 1–8. https://doi.org/10.1109/FUZZ-IEEE.2018.8491501
158. Murray BJ, Islam MA, Pinar AJ, Anderson DT, Scott GJ, Havens TC, Keller JM. Explainable AI for the Choquet Integral. IEEE Trans Emerg Topics in Comput Intell. 2020. https://doi.org/10.1109/TETCI.2020.3005682.
159. Murrell S, Plant TR. A survey of tools for the validation and verification of knowledge-based systems: 1985–1995. Decis Support Syst. 1997;21(4):307–23. https://doi.org/10.1016/S0167-9236(97)00047-X.
160. Mynuddin M, Gao W. Distributed predictive cruise control based on reinforcement learning and validation on microscopic traffic simulation. IET Intel Transport Syst. 2020;14(5):270–7. https://doi.org/10.1049/iet-its.2019.0404.
161. Nassar M, Salah K, Rehman MH, Svetinovic D. Blockchain for explainable and trustworthy artificial intelligence. WIREs Data Mining Knowl Disc. 2020;10(1):e1340. https://doi.org/10.1002/widm.1340.
162. Niazi M. A, Siddique Q, Hussain A, Kolberg M. (2010). Verification & validation of an agent-based forest fire simulation model. Proceedings of the 2010 Spring Simulation Multiconference, 1–8. https://doi.org/10.1145/1878537.1878539
163. Nourani CF. Multi-agent object level AI validation and verification. ACM SIGSOFT Softw Eng Notes. 1996;21(1):70–2. https://doi.org/10.1145/381790.381802.
164. O'Keefe RM, Balci O, Smith EP. Validating expert system performance. IEEE Expert. 1987;2(4):81–90. https://doi.org/10.1109/MEX.1987.5006538.
165. On Artificial Intelligence—A European approach to excellence and trust. (2020). European Commision.
166. Onoyama T, Tsuruta S. Validation method for intelligent systems. J Exp Theor Artif Intell. 2000;12(4):461–72. https://doi.org/10.1080/095281300454838.
167. Pawar U, O'Shea D, Rea S, O'Reilly R. (2020). Explainable AI in Healthcare. 2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), 1–2. https://doi.org/10.1109/CyberSA49311.2020.9139655
168. Payrovnaziri SN, Chen Z, Rengifo-Moreno P, Miller T, Bian J, Chen JH, Liu X, He Z. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. J Am Med Inform Assoc. 2020;27(7):1173–85. https://doi.org/10.1093/jamia/ocaa053.
169. Pèpe G, Perbost R, Courcambeck J, Jouanna P. Prediction of molecular crystal structures using a genetic algorithm: validation by GenMolTM on energetic compounds. J Cryst Growth. 2009;311(13):3498–510. https://doi.org/10.1016/j.jcrysgro.2009.04.002.
170. Peppler RA, Long CN, Sisterson DL, Turner DD, Bahrmann CP, Christensen SW, Doty KJ, Eagan RC, Halter TD, Iveyh MD, Keck NN, Kehoe KE, Liljegren JC, Macduff MC, Mather JH, McCord RA, Monroe JW, Moore ST, Nitschke KL, Wagener R. An overview of ARM program climate research facility data quality assurance. Open Atmos Sci J. 2008;2(1):192–216. https://doi.org/10.2174/1874282300802010192.
171. Pitchforth, J. (2013). A proposed validation framework for expert elicited Bayesian Networks. Expert Systems with Applications, 6.
172. Pocius R, Neal L, Fern A. Strategic tasks for explainable reinforcement learning. Proc AAAI Conf Artif Intell. 2019;33:10007–8. https://doi.org/10.1609/aaai.v33i01.330110007.
173. Preece AD, Shinghal R, Batarekh A. Verifying expert systems: a logical framework and a practical tool. Expert Syst Appl. 1992;5(3–4):421–36. https://doi.org/10.1016/0957-4174(92)90026-O.
174. Prentzas, N, Nicolaides, A, Kyriacou, E, Kakas, A, Pattichis, C. (2019). Integrating Machine Learning with Symbolic Reasoning to Build an Explainable AI Model for Stroke Prediction. 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), 817–821. https://doi.org/10.1109/BIBE.2019.00152
175. Puiutta E, Veith E. M. (2020). Explainable Reinforcement Learning: A Survey. ArXiv:2005.06247 [Cs, Stat].n
176. Putzer, H. J, Wozniak E. (2020). A Structured Approach to Trustworthy Autonomous/Cognitive Systems. ArXiv:2002.08210 [Cs].n
177. Pynadath DV. Transparency communication for machine learning. In human-automation interaction human and machine learning. Berlin: Springer International Publishing; 2018.
178. Qiu S, Liu Q, Zhou S, Wu C. Review of artificial intelligence adversarial attack and defense technologies. Appl Sci. 2019;9(5):909. https://doi.org/10.3390/app9050909.
179. Ragot M, Martin N, Cojean S. (2020). AI-generated vs. Human Artworks. A Perception Bias Towards Artificial Intelligence? Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, 1–10. https://doi.org/10.1145/3334480.3382892
180. Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson, B, Smith-Loud, J, Theron D, Barnes P. (2020). Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. ArXiv:2001.00973 [Cs].n
181. Raymond P, Yoav S, Erik B, Jack C, John E, Barbara G, Terah L, James M, Juan C N, Saurabh M. (2020). Artificial Intelligence Index 2019 Annual report [Artificial Intelligence Index Annual Report]. Stanford University Human AI. Available at: https://hai.stanford.edu/sites/default/files/ai_index_2019_report.pdf
182. Ren H, Chandrasekar S. K, Murugesan A. (2019). Using Quantifier Elimination to Enhance the Safety Assurance of Deep Neural Networks. ArXiv:1909.09142 [Cs, Stat].n

183. Rossi F. (2018). Building Trust in Artificial Intelligence. Undefined. /paper/Building-Trust-in-Artificial-Intelligence-Rossi/e7a84026ac8806bd377b5b491c57096083bbbb18
184. Rotman, N. H, Schapira M, Tamar, A. (2020). Online Safety Assurance for Deep Reinforcement Learning. ArXiv:2010.03625 [Cs].
185. Rovcanin M, De Poorter E, van den Akker D, Moerman I, Demeester P, Blondia C. Experimental validation of a reinforcement learning based approach for a service-wise optimisation of heterogeneous wireless sensor networks. Wireless Netw. 2015;21(3):931–48. https://doi.org/10.1007/s11276-014-0817-8.
186. Ruan Y, Zhang P, Alfantoukh L, Durresi A. Measurement Theory-Based Trust Management Framework for Online Social Communities. ACM Transactions on Internet Technology. 2017;17(2):1–24. https://doi.org/10.1145/3015771.
187. Ruan W, Huang X, Kwiatkowska M (2018). Reachability Analysis of Deep Neural Networks with Provable Guarantees. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 2651–2659. https://doi.org/10.24963/ijcai.2018/368
188. Sarathy N, Alsawwaf M, Chaczko Z. (2020). Investigation of an Innovative Approach for Identifying Human Face-Profile Using Explainable Artificial Intelligence. 2020 IEEE 18th International Symposium on Intelligent Systems and Informatics (SISY), 155–160. https://doi.org/10.1109/SISY50555.2020.9217095
189. Sargent RG. Verification and validation of simulation models. J Simul. 2013;7(1):12–24. https://doi.org/10.1057/jos.2012.20.
190. Sargent RG (1984). A tutorial on verification and validation of simulation models. Proceedings of the 16th Conference on Winter Simulation, 114–121.
191. Sargent RG (2004). Validation and Verification of Simulation Models. Proceedings of the 2004 Winter Simulation Conference, 2004, 1, 13–24. https://doi.org/https://doi.org/10.1109/WSC.2004.1371298
192. Sargent RG. (2010). Verification and validation of simulation models. Proceedings of the 2010 Winter Simulation Conference, 166–183. https://doi.org/10.1109/WSC.2010.5679166
193. Schlegel U, Arnout H, El-Assady M, Oelke D, Keim D. A. (2019). Towards a Rigorous Evaluation of XAI Methods on Time Series. ArXiv:1909.07082 [Cs].
194. Schumann J, Gupta P, Liu Y. Application of neural networks in high assurance systems: a survey. In: Schumann J, Liu Y, editors. Applications of neural networks in high assurance systems, vol. 268. Berlin: Springer; 2010. p. 1–19; https://doi.org/10.1007/978-3-642-10690-3_1.
195. Schumann J, Gupta, P, Nelson S. (2003). On verification validation of neural network based controllers.
196. Sequeira P, Gervasio M. Interestingness elements for explainable reinforcement learning: understanding agents' capabilities and limitations. Artif Intell. 2020;288:103367. https://doi.org/10.1016/j.artint.2020.103367.
197. Sileno G, Boer A, van Engers T. (2018). The Role of Normware in Trustworthy and Explainable AI.ArXiv:1812.02471 [Cs].
198. Singer E, Thurn DRV, Miller ER. Confidentiality assurances and response: a quantitative review of the experimental literature. Public Opin Q. 1995;59(1):66. https://doi.org/10.1086/269458.
199. Sivamani KS, Sahay R, Gamal AE. Non-intrusive detection of adversarial deep learning attacks via observer networks. IEEE Lett Comput Soc. 2020;3(1):25–8. https://doi.org/10.1109/LOCS.2020.2990897.
200. Spada M. R, Vincentini A. (2019). Trustworthy AI for 5G: Telco Experience and Impact in the 5G ESSENCE. In J. MacIntyre, I. Maglogiannis, L. Iliadis, E. Pimenidis (Eds.), Artificial Intelligence Applications and Innovations (pp. 103–110). Springer International Publishing. https://doi.org/10.1007/978-3-030-19909-8_9
201. Spinner T, Schlegel U, Schafer H, El-Assady M. explAIner: a visual analytics framework for interactive and explainable machine learning. IEEE Trans Visual Comput Graph. 2019. https://doi.org/10.1109/TVCG.2019.2934629.
202. Srivastava B, Rossi, F. (2019). Towards Composable Bias Rating of AI Services.ArXiv:1808.00089 [Cs].
203. Stock P, Cisse M. ConvNets and Imagenet beyond accuracy: understanding mistakes and uncovering biases. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors. Computer Vision – ECCV 2018, vol. 11210. Berlin: Springer; 2018. p. 504–19.
204. Suen CY, Grogono PD, Shinghal R, Coallier F. Verifying, validating, and measuring the performance of expert systems. Expert Syst Appl. 1990;1(2):93–102. https://doi.org/10.1016/0957-4174(90)90019-Q.
205. Sun SC, Guo, W. (2020). Approximate Symbolic Explanation for Neural Network Enabled Water-Filling Power Allocation. 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), 1–4. https://doi.org/10.1109/VTC2020-Spring48590.2020.9129447
206. Tadj C. Dynamic verification of an object-rule knowledge base using colored petri Nets. System Cybern Inf. 2005;4(3):9.
207. Tan R, Khan N, Guan L. Locality guided neural networks for explainable artificial intelligence. International Joint Conference on Neural Networks. 2020;2020:1–8. https://doi.org/10.1109/IJCNN48605.2020.9207559.
208. Tao C, Gao J, Wang T. Testing and quality validation for AI software-perspectives issues, and practices. IEEE Access. 2019;7:12.
209. Tao, J, Xiong, Y, Zhao, S, Xu, Y, Lin, J, Wu, R, Fan, C. (2020). XAI-Driven Explainable Multi-view Game Cheating Detection. 2020 IEEE Conference on Games (CoG), 144–151. https://doi.org/10.1109/CoG47356.2020.9231843
210. Taylor BJ, Darrah MA (2005). Rule extraction as a formal method for the verification and validation of neural networks. Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005, 5, 2915–2920. https://doi.org/10.1109/IJCNN.2005.1556388
211. Taylor, Brian J. (Ed.). (2006). Methods and Procedures for the Verification and Validation of Artificial Neural Networks. Springer US. https://doi.org/10.1007/0-387-29485-6
212. Taylor BJ, Darrah MA, Moats CD (2003). Verification and validation of neural networks: A sampling of research in progress (K. L. Priddy P. J. Angeline, Eds.; p. 8). https://doi.org/10.1117/12.487527
213. Taylor, E, Shekhar, S, Taylor, G. W. (2020). Response Time Analysis for Explainability of Visual Processing in CNNs. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 1555–1558. https://doi.org/10.1109/CVPRW50498.2020.00199
214. Thomas JD, Sycara K. (1999). The Importance of Simplicity and Validation in Genetic Programming for Data Mining in Financial Data. AAAI Technical Report, 5.

215. Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): towards medical XAI. IEEE Trans Neural Netw Learn Syst. 2020. https://doi.org/10.1109/TNNLS.2020.3027314.
216. Toreini E, Aitken M, Coopamootoo K, Elliott K, Zelaya CG (2020). The relationship between trust in AI and trustworthy machine learning technologies. 12.
217. Toreini E, Aitken M, Coopamootoo KPL, Elliott K, Zelaya VG, Missier P, Ng M, van Moorsel A (2020). Technologies for Trustworthy Machine Learning: A Survey in a Socio-Technical Context.ArXiv:2007.08911 [Cs, Stat].
218. Tsai W-T, Vishnuvajjala R, Zhang D. Verification and validation of knowledge-based systems. IEEE Trans Knowl Data Eng. 1999;11(1):11.
219. Turing A. Computing machinery and intelligence. Mind. 1950;59(236):433–60.
220. Uslu S, Kaur D, Rivera SJ, Durresi A, Babbar-Sebens M. Trust-Based game-theoretical decision making for food-energy-water management. In: Barolli L, Hellinckx P, Enokido T, editors. Advances on broad-band wireless computing, communication and applications, vol. 97. Berlin: Springer; 2020. p. 125–36; https://doi.org/10.1007/978-3-030-33506-9_12.
221. Uslu S, Kaur D, Rivera SJ, Durresi A, Babbar-Sebens M. Trust-based decision making for food-energy-water actors. In: Barolli L, Amato F, Moscato F, Enokido T, Takizawa M, editors. Advanced information networking and applications. Berlin: Springer International Publishing; 2020. p. 591–602.
222. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS ONE. 2019;14(11):e0224365. https://doi.org/10.1371/journal.pone.0224365.
223. Validation of Machine Learning Models: Challenges and Alternatives. (2017). protiviti.
224. Varshney KR. Trustworthy machine learning and artificial intelligence. XRDS. 2019;25(3):26–9. https://doi.org/10.1145/3313109.
225. Varshney KR (2020). On Mismatched Detection and Safe, Trustworthy Machine Learning. 2020 54th Annual Conference on Information Sciences and Systems (CISS), 1–4. https://doi.org/10.1109/CISS48834.2020.1570627767
226. Veeramachaneni K, Arnaldo I, Korrapati V, Bassias C, Li K. (2016). AI2: Training a Big Data Machine to Defend. 2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS), 49–54. https://doi.org/10.1109/BigDataSecurity-HPSC-IDS.2016.79
227. Vinze AS, Vogel DR, Nunamaker JF. Performance evaluation of a knowledge-based system. Inf Manag. 1991;21(4):225–35. https://doi.org/10.1016/0378-7206(91)90068-D.
228. Volz V, Majchrzak K, Preuss M. (2018). A Social Science-based Approach to Explanations for (Game) AI. 2018 IEEE Conference on Computational Intelligence and Games (CIG), 1–2. https://doi.org/10.1109/CIG.2018.8490361
229. Wang D, Yang Q, Abdul A, Lim BY (2019). Designing Theory-Driven User-Centric Explainable AI. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19, 1–15. https://doi.org/10.1145/3290605.3300831
230. Wei S, Zou Y, Zhang T, Zhang X, Wang W. Design and experimental validation of a cooperative adaptive cruise control system based on supervised reinforcement learning. Appl Sci. 2018;22:1014.
231. Welch ML, McIntosh C, Traverso A, Wee L, Purdie TG, Dekker A, Haibe-Kains B, Jaffray DA. External validation and transfer learning of convolutional neural networks for computed tomography dental artifact classification. Phys Med Biol. 2020;65(3):035017. https://doi.org/10.1088/1361-6560/ab63ba.
232. Wells SA (1993). The VIVA Method: A Life-cycle Independent Approach to KBS Validation. AAAI Technical Report WS-93–05, 5.
233. Wickramage N (2016). Quality assurance for data science: Making data science more scientific through engaging scientific method. 2016 Future Technologies Conference (FTC). https://doi.org/10.1109/FTC.2016.7821627
234. Wieringa M (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 1–18. https://doi.org/10.1145/3351095.3372833
235. Wing JM (2020). Trustworthy AI. ArXiv:2002.06276[Cs].
236. Winkel DJ. Validation of a fully automated liver segmentation algorithm using multi-scale deep reinforcement learning and comparison versus manual segmentation. Eur J Radiol. 2020;7:108918.
237. Winkler T, Rinner B. (2010). User-Based Attestation for Trustworthy Visual Sensor Networks. 2010 IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing, 74–81. https://doi.org/10.1109/SUTC.2010.20
238. Wu C-H, Lee S-J. KJ3—A tool assisting formal validation of knowledge-based systems. Int J Hum Comput Stud. 2002;56(5):495–524. https://doi.org/10.1006/ijhc.2002.1007.
239. Xiao Y, Pun C-M, Liu B. Adversarial example generation with adaptive gradient search for single and ensemble deep neural network. Inf Sci. 2020;528:147–67. https://doi.org/10.1016/j.ins.2020.04.022.
240. Xu W, Evans D, Qi Y. (2018). Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. Proceedings 2018 Network and Distributed System Security Symposium. https://doi.org/10.14722/ndss.2018.23198
241. Yilmaz L. Validation and verification of social processes within agent-based computational organization models. Comput Math Organ Theory. 2006;12(4):283–312. https://doi.org/10.1007/s10588-006-8873-y.
242. Yoon J, Kim K, Jang J. (2019). Propagated Perturbation of Adversarial Attack for well-known CNNs: Empirical Study and its Explanation. 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 4226–4234. https://doi.org/10.1109/ICCVW.2019.00520
243. Zaidi AK, Levis AH. Validation and verification of decision making rules. Automatica. 1997;33(2):155–69. https://doi.org/10.1016/S0005-1098(96)00165-3.
244. Zeigler BP, Nutaro JJ. Towards a framework for more robust validation and verification of simulation models for systems of systems. J Def Model Simul . 2016;13(1):3–16. https://doi.org/10.1177/1548512914568657.
245. Zhou J, Chen F (2019). Towards Trustworthy Human-AI Teaming under Uncertainty. 5.
246. Zhu, H, Xiong, Z, Magill, S, Jagannathan, S. (2019). An inductive synthesis framework for verifiable reinforcement learning. Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation, 686–701. https://doi.org/10.1145/3314221.3314638

247. Zhu J, Liapis A, Risi S, Bidarra R, Youngblood GM (2018). Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation. 2018 IEEE Conference on Computational Intelligence and Games (CIG), 1–8. https://doi.org/10.1109/CIG.2018.8490433
248. Zlatareva NP (1998). Knowledge Refinement during Developmental and Field Validation of Expert Systems. 6.
249. Zlatareva N, Preece A. State of the art in automated validation of knowledge-based systems. Expert Syst Appl. 1994;7(2):151–67. https://doi.org/10.1016/0957-4174(94)90034-5.

Publisher's Note