A Survey on Artificial Intelligence Assurance

Feras A. Batarseh; Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University (Virginia Tech), Arlington, VA 22203 (*Corresponding author) batarseh@vt.edu

Laura Freeman; Department of Statistics, Virginia Polytechnic Institute and State University (Virginia Tech), Arlington, VA 22203 laura.freeman@vt.edu

Chih-Hao Huang; College of Science, George Mason University, Fairfax 22030 chuang21@gmu.edu

Abstract – Artificial Intelligence (AI) algorithms are increasingly providing decision making and operational support across multiple domains. AI includes a wide library of algorithms for different problems. One important notion for the adoption of AI algorithms into operational decision process 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 – 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

1. 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.

1.1 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” (DoD, 1995).

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, section 2 provides a historical perspective as well as the entire assurance landscape, section 3 includes an exhaustive list of papers relevant to AI assurance (as well as the scoring system), section 4 presents overall insights and discussions of the survey, and lastly, section 5 presents conclusions.

1.2 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.

2. 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.

2.1 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 (Turing 1950). 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
Comparing predictions (or classifications) to actual outputs is one way of evaluating that the results of an algorithm match what the real world created.

There were a dominating number of validation and verification methods in the seventies, eighties, and nineties for two forms of intelligence, knowledge-based systems (i.e., expert systems) and simulation systems (majorly for defense and military applications). One of the first times where AI turned towards data-driven methods was apparent in 1996 at the Third International Math and Science Study (TIMSS), which focused on quality assurance in data collection (Martin and Mullis, 1996). Data from Forty-five countries were included in the analysis. In a very deliberate process, the data collectors were faced with challenges relevant to the internationalization of data. For example, data from Indonesia had errors in translation; data collection processes were different in Korea, Germany, and Kuwait than the standard process due to funding and timing issues. Such real-world issues in data collection certainly pose a challenge to the assurance of statistical learning AI that require addressing.

In the 1990s, AI testing and assurance were majorly inspired by the big research archive of testing of software (i.e., within software engineering) (Batarseh et al., 2020). However, a slim amount of literature explored algorithms such as genetic algorithms (Jones et al., 1997), 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 (Berndt et al., 2001), or algorithms-specific assurance such as Crisp Clustering for k-means clustering (Halkidi et al., 2001).

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 Figure 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 Figure 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 (2.2) provides further details on the state-of-the-art, and section 3 presents an exhaustive review of all AI assurance methods found under the predefined search criteria.

2.2 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” (Taylor, 2006), 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 (Batarseh & Gonzalez, 2013). 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 (Massoli et al., 2021). 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 (Kulkarni et al., 2020).

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 (Tsai et al., 1999), (Batarseh & Gonzalez, 2015), and (Onoyama & Tsuruta, 2000), other “more recent” methods were focused on one AI subarea and one domain. For instance, in Mason et al. (2017), assurance was applied to reinforcement learning methods for safety-critical systems. Precentzas et al. (2019) presented an assurance method for machine learning as its applied to stroke predictions, similar to Pawar’s et al.’s (2020) XAI for healthcare framework. Pepe et al. (2009), and Chittajallu et al.’s (2019) 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 (Magazzeni, 2017), 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) (Schlegel et al., 2019). Few papers without substantial formal methods were found for Fair AI, Safe AI (Everitt, 2018), Transparent AI (Behnoush & Nasraoui, 2018), or Trustworthy AI (Aitken et al., 2016); but XAI (Hagras, 2018) has been central (as the previous figures in this paper also suggest). For instance, in Lee et al. (2019), 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. (2019) presented a model for XAI that is tailored for road traffic forecasting, and Guo (2020) presented the same, albeit for 5G and wireless networks (Spada & Vincentini, 2019). Similarly, Kuppa and Le-Khac (2020) presented a method focused on Cyber Security using gradient maps and bots. Go & Lee (2018) presented an AI assurance method for trustworthiness of security systems. Lastly, Guo (2020) 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 (Nourani et al., 2016). The challenges of AI
algorithms and their assurance is evident and consistent across many of the manuscripts, such as in Janssen and Kuk’s (2016) study of the limitations of AI for government, on the other hand, Batarseh et al. (2017) 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 (Jorge et al., 2018), (Bruno et al., 2017), and (Laat, 2017). Other less “time-constrained” activities such as decisions at organizations (Ruan, 2017) and time series decision support systems could utilize slower methods such as genetic algorithms (Thomas & Sycara, 1999), 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” (Thomas & Sycara, 1999) when addressing the validation part of their work. A recent publication by Raji et al. (2020) 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 (Srivastava & Rossi, 2019) (Varshney, 2020). 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.

3. The review and scoring of 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 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 1 presents a list of all reviewed manuscripts and their detailed scores (for the ten metrics) by ranking category. The papers, data, dashboard, and lists are on a public GitHub repository: [https://github.com/ferasbatarseh/AI-Assurance-Review](https://github.com/ferasbatarseh/AI-Assurance-Review)

| Year | First Author's Last Name and Citation | Publishing Venue | AI Subarea | Total Score |
|------|--------------------------------------|-----------------|------------|-------------|
| 2020 | D’Alterio (D’Alterio et al., 2020)   | FUZZ-IEEE       | XAI        | 10          |
| 2019 | Tao (C. Tao et al., 2019)            | IEEE Access     | Generic    | 10          |
| 2020 | Anderson (A. Anderson et al., 2020) | ACM TIIS       | RL         | 9           |
| 2020 | Birkenbihl (Birkenbihl, 2020)       | EPMA            | ML         | 9           |
| 2020 | Checco (Checco et al., 2020)        | JAIR            | DS         | 9           |
| 2020 | Chen (H.-Y. Chen & Lee, 2020)       | IEEE Access     | XAI        | 9           |
| 2020 | Cluzeau (Cluzeau et al., 2020)      | EASA            | DL         | 9           |
| 2019 | Kaur (Kaur et al., 2019)            | WAINA           | XAI        | 9           |
| Year | Authors (Ref. Year) | Journal / Conference | Volume | Page |
|------|---------------------|----------------------|--------|------|
| 2020 | Kulkarni (Kulkarni et al., 2020) | Academic Press | DS | 9 |
| 2020 | Kuppa (Kuppa & Le-Khac, 2020) | IEEE IJCNN | XAI | 9 |
| 2020 | Kuzlu (Kuzlu et al., 2020) | IEEE Access | XAI | 9 |
| 2021 | Massoli (Massoli et al., 2021) | CVIU | DL | 9 |
| 2020 | Spinner (Spinner et al., 2019) | IEEE TVCG | XAI | 9 |
| 2016 | Veeramachaneni (Veeramachaneni et al., 2016) | IEEE HPSC | DS | 9 |
| 2018 | Wei (Wei et al., 2018) | AS | RL | 9 |
| 2020 | Winkel (Winkel, 2020) | EJR | RL | 9 |
| 2014 | Ali (Ali & Schmid, 2014) | GISci | DS | 8 |
| 2018 | Alves (Alves et al., 2018) | NASA ARIAS | ABS | 8 |
| 2019 | Batarseh (Batarseh & Kulkarni, 2019) | EDML | DS | 8 |
| 2016 | Gao (Gao et al., 2016) | SEKE | DS | 8 |
| 2020 | Gardiner (Gardiner et al., 2020) | Nature Sci Rep | ML | 8 |
| 2016 | Gulshan (Gulshan et al., 2016) | JAMA | CV | 8 |
| 2020 | Guo (Guo, 2020a) | IEEE ICCVW | XAI | 8 |
| 2020 | Han (Han et al., 2020) | IET JoE | XAI | 8 |
| 2016 | Heaney (Heaney et al., 2016) | OD | GA | 8 |
| 2019 | Huber (Huber, 2019) | KI AAI | RL | 8 |
| 2019 | Keneni (Keneni et al., 2019) | IEEE Access | XAI | 8 |
| 2020 | Kohlbrenner (Kohlbrenner et al., 2020) | IEEE IJCNN | XAI | 8 |
| 2019 | Maloca (Maloca et al., 2019) | PLoS ONE | DL | 8 |
| 2020 | Malolan (Malolan et al., 2020) | IEEE ICICT | XAI | 8 |
| 2020 | Payrovnaziri (Payrovnaziri et al., 2020) | JAMIA | ML | 8 |
| 2008 | Peppler (Peppler et al., 2008) | OASJ | DS | 8 |
| 2020 | Sequeira (Sequeira & Gervasio, 2020) | SciDir AI | RL | 8 |
| 2020 | Sivamani (Sivamani et al., 2020) | IEEE LCS | DL | 8 |
| 2020 | Tan (Tan et al., 2020) | IEEE IJCNN | XAI | 8 |
| 2020 | Tao (J. Tao et al., 2020) | IEEE CoG | XAI | 8 |
| 2020 | Welch (Welch et al., 2020) | PhysMedBiol | DL | 8 |
| 2020 | Xiao (Xiao et al., 2020) | IS | DL | 8 |
| 2016 | Aitken (Aitken, 2016) | UC | ABS | 7 |
| 2019 | Barredo-Arrieta (Barredo-Arrieta et al., 2019) | IEEE ITSC | XAI | 7 |
| 2013 | Batarseh (Batarseh & Gonzalez, 2013) | IEEE TSMCS | KBS | 7 |
| 2001 | Berndt (Berndt et al., 2001) | COMP | DS | 7 |
| 2010 | Bone (Bone & Dragićević, 2010) | CEUS | RL | 7 |
| 2016 | Celis (Celis et al., 2016) | PrePrint | ML | 7 |
| Year | Authors (Title of Work) | Conference/Journal | Area | Pages |
|------|-------------------------|--------------------|------|-------|
| 2019 | Chittajallu (Chittajallu et al., 2019) | IEEE ISBI | XAI | 7 |
| 2018 | Elsayed (Elsayed et al., 2018) | NIPS | CV | 7 |
| 2019 | Ferreyra (Ferreyra et al., 2019) | FUZZ-IEEE | XAI | 7 |
| 2006 | Forster (Forster, 2006) | Uni of South Africa | AGI | 7 |
| 1985 | Ginsberg (Ginsberg & Weiss, 2001) | IJCAI | KBS | 7 |
| 2018 | Go (Go & Lee, 2018) | ACM CCS | DL | 7 |
| 2020 | Halliwell (Halliwell & Lecue, 2020) | PrePrint | DL | 7 |
| 2015 | He (C. He et al., 2015) | MPE | GA | 7 |
| 2020 | Heuer (Heuer & Breiter, 2020) | ACM UMAP | ML | 7 |
| 2016 | Jiang (Jiang & Li, 2016) | PMLR | RL | 7 |
| 2020 | Kaur (Kaur et al., 2020) | AINA | XAI | 7 |
| 2016 | Kianifar (Kianifar, 2016) | SC | GA | 7 |
| 2019 | Lee (J. ha Lee et al., 2019) | IEEE ICTC | XAI | 7 |
| 2017 | Liang (Liang et al., 2017) | MILCOM | DS | 7 |
| 2020 | Mackowiak (Mackowiak et al., 2020) | PrePrint | CV | 7 |
| 2018 | Mason (Mason et al., 2018) | AHIM | RL | 7 |
| 2018 | Murray (B. Murray et al., 2018) | FUZZ-IEEE | XAI | 7 |
| 2019 | Naqa (El Naqa et al., 2019) | MedPhys | ML | 7 |
| 2019 | Prentzas (Prentzas et al., 2019) | IEEE BIBE | XAI | 7 |
| 2018 | Pynadath (Pynadath, 2018) | Springer HCIS | ML | 7 |
| 2020 | Ragot (Ragot et al., 2020) | CHI | ML | 7 |
| 2020 | Rotman (Rotman et al., 2020) | PrePrint | RL | 7 |
| 2015 | Rovcanin (Rovcanin et al., 2015) | WN | RL | 7 |
| 2020 | Sarathy (Sarathy et al., 2020) | IEEE SISY | XAI | 7 |
| 2018 | Stock (Stock & Cisse, 2018) | ECCV | CV | 7 |
| 2009 | Tadj (Tadj, 2005) | SCI | KBS | 7 |
| 1999 | Thomas (Thomas & Sycara, 1999) | AAAI | GA | 7 |
| 2020 | Uslu (Uslu et al., 2020a) | AINA | XAI | 7 |
| 2018 | Xu (Xu et al., 2018) | PrePrint | DL | 7 |
| 2019 | Bellamy (Bellamy et al., 2019) | IBM JRD | XAI | 6 |
| 2019 | Beyret (Beyret et al., 2019) | IEEE IROS | RL | 6 |
| 2018 | Cao (Cao et al., 2019) | JAIHC | ML | 6 |
| 2020 | Cruz (Cruz et al., 2020) | PrePrint | RL | 6 |
| 2001 | Halkidi (Halkidi et al., 2001) | JIIS | ML | 6 |
| 2020 | He (Y. He et al., 2020) | PrePrint | RL | 6 |
| 2020 | Islam (Islam et al., 2019) | IEEE TFS | XAI | 6 |
| 2005 | Liu (F. Liu & Yang, 2005) | AI2005 | DL | 6 |
| Year | Author(s) (Title of paper, Year) | Journal | Conference | Volume |
|------|----------------------------------|---------|------------|--------|
| 2019 | Madumal (Madumal et al., 2019)  | PrePrint| RL         | 6      |
| 1996 | Martin (Martin et al., 1996)    | ERIC    | DS         | 6      |
| 2007 | Martín-Guerrero (Martín-Guerrero et al., 2007) | AJCAI    | RL         | 6      |
| 2000 | Mosqueira-Rey (Mosqueira-Rey & Moret-Bonillo, 2000) | ESA      | KBS        | 6      |
| 2020 | Mynuddin (Mynuddin & Gao, 2020) | IETITS  | RL         | 6      |
| 2020 | Puiutta (Puiutta & Veith, 2020) | CD-MAKE | RL         | 6      |
| 2018 | Ruan (Ruan et al., 2018)        | IJCAI   | DL         | 6      |
| 2019 | Schlegel (Schlegel et al., 2019) | IEEE ICCVW | XAI     | 6      |
| 2020 | Toreini (Toreini et al., 2020)  | ACM FAT | ML         | 6      |
| 2020 | Toreini (Toreini et al., 2020)  | PrePrint| ML         | 6      |
| 2019 | Vabalas (Vabalas et al., 2019)  | PLoS ONE| ML         | 6      |
| 2010 | Winkler (Winkler & Rinner, 2010) | IEEE SUTC | CV     | 6      |
| 2002 | Wu (Wu & Lee, 2002)             | IJHCS   | KBS        | 6      |
| 2019 | Zhu (H. Zhu et al., 2019)       | ACM PLDI| RL         | 6      |
| 1992 | Andert (Andert, 1992)           | IJM     | KBS        | 5      |
| 2018 | Antunes (Antunes et al., 2018)  | IEEE DSN-W | ML   | 5      |
| 1989 | Becker (Becker et al., 1989)    | NASA    | KBS        | 5      |
| 2019 | Chen (T. Chen et al., 2019)     | CS      | RL         | 5      |
| 2019 | Cruz (Cruz et al., 2019)        | AI 2019 AAI | RL   | 5      |
| 2020 | Diallo (Diallo et al., 2020)    | IEEE ACSOS-C | XAI | 5      |
| 2010 | Dong (Dong et al., 2010)        | IEEE ICWIIAT | GA     | 5      |
| 2019 | Dupuis (Dupuis & Verheij, 2019) | UoG     | XAI        | 5      |
| 2015 | Goodfellow (Goodfellow et al., 2015) | PrePrint | ML      | 5      |
| 2020 | Guo (Guo, 2020b)                | IEEE CM  | XAI        | 5      |
| 2020 | Haverinen (Haverinen, 2020)    | Un of Jyväskylä | XAI | 5      |
| 1997 | Jones (Jones et al., 1997)      | JMB     | GA         | 5      |
| 2019 | Joo (Joo & Kim, 2019)           | IEEE CoG| RL         | 5      |
| 2020 | Katell (Katell et al., 2020)    | ACM FAT | XAI        | 5      |
| 2007 | Knauf (Rainer Knauf et al., 2007) | IEEE TSMC | KBS | 5      |
| 1995 | Lockwood (Lockwood & Chen, 1995) | AES      | KBS        | 5      |
| 2000 | Marcos (Marcos et al., 2000)    | IEE Proc | KBS        | 5      |
| 2017 | Mason (Mason et al., 2017b)     | WhiteRose| RL         | 5      |
| 1988 | Morell (Morell, 1988)           | IEA/AIE | KBS        | 5      |
| 2020 | Murray (B. J. Murray et al., 2020) | IEEE TETCI | XAI | 5      |
| 2010 | Niazi (Niazi et al., 2010)      | SpringSim | ABS      | 5      |
| 2000 | Onoyama (Onoyama & Tsuruta, 2000) | JETAI    | KBS        | 5      |
| Year | Authors                                  | Journal/Conference | Section | Pages |
|------|------------------------------------------|--------------------|---------|-------|
| 2019 | Ren (Ren et al., 2019)                   | PrePrint DL        |         | 5     |
| 2013 | Sargent (Robert G. Sargent, 2013)        | JoS ABS DL         |         | 5     |
| 2003 | Schumann (Schumann et al., 2003)        | EANN DL            |         | 5     |
| 1995 | Singer (Singer et al., 1995)            | POQ DS             |         | 5     |
| 2019 | Srivastava (Srivastava & Rossi, 2019)   | AAAI AIES NLP      |         | 5     |
| 2006 | Taylor (Brian J. Taylor, 2006)          | Springer DL        |         | 5     |
| 2020 | Taylor (E. Taylor et al., 2020)         | IEEE CVPRW XAI    |         | 5     |
| 2020 | Tjoa (Tjoa & Guan, 2020)                | IEEE TNNLS ML      |         | 5     |
| 2020 | Uslu (Uslu et al., 2020b)               | BWCCA XAI         |         | 5     |
| 2020 | Varshney (Varshney, 2020)               | IEEE CISS ML      |         | 5     |
| 2018 | Volz (Volz et al., 2018)                | IEEE CIG XAI      |         | 5     |
| 2020 | Wieringa (Wieringa, 2020)               | ACM FAT XAI      |         | 5     |
| 2020 | Wing (Wing, 2020)                       | PrePrint ML       |         | 5     |
| 2019 | Yoon (Yoon et al., 2019)                | IEEE ICCVW XAI   |         | 5     |
| 2019 | Zhou (Zhou & Chen, 2019)                | IJCAI XAI ML      |         | 5     |
| 1994 | Zlatareva (N. Zlatareva & Preece, 1994) | ESA KBS           |         | 5     |
| 2018 | Al Now (Algorithmic Accountability Policy Tooklit, 2018) | AI Now XAI |         | 4     |
| 2015 | Arifin (Arifin & Madey, 2015)           | Springer ABS      |         | 4     |
| 2015 | Batarseh (Batarseh & Gonzalez, 2015)    | AIR KBS           |         | 4     |
| 2007 | Brancovici (Brancovici, 2007)           | IEEE CEC XAI     |         | 4     |
| 1987 | Castore (Castore, 1987)                 | NASA STI KBS      |         | 4     |
| 2013 | Cohen (Cohen et al., 2013)              | EternalS NLP     |         | 4     |
| 2020 | Das (Das & Rad, 2020)                   | PrePrint XAI      |         | 4     |
| 2013 | David (David, 2013)                     | UCS ABS           |         | 4     |
| 2018 | Došilović (Došilović et al., 2018)     | MIPRO ML         |         | 4     |
| 2000 | Edwards (Edwards, 2000)                 | Oxford DS        |         | 4     |
| 2018 | EY (Assurance in the Age of AI, 2018)   | EY ML            |         | 4     |
| 2019 | Guidotti (Guidotti et al., 2019)        | ACM CS XAI       |         | 4     |
| 2018 | Jilk (Jilk, 2018)                       | PrePrint ABS     |         | 4     |
| 2017 | Leibovici (Leibovici et al., 2017)      | ISPRS Int J. Geo-Inf DS |     | 4     |
| 2020 | Li (Li et al., 2020)                    | IEEE TKDE XAI    |         | 4     |
| 2019 | Mehrabi (Mehrabj et al., 2019)          | PrePrint ML       |         | 4     |
| 2019 | Meskauskas (Meskauskas et al., 2020)    | FUZZ-IEEE XAI    |         | 4     |
| 1998 | Miller (Miller, 1998)                   | MS GA            |         | 4     |
| 2019 | Nassar (Nassar et al., 2020)            | WIREs DMKD XAI   |         | 4     |
| 1992 | Preece (Preece et al., 1992)            | ESA KBS          |         | 4     |
| Year | Author(s) | Conference/Book | Journal/Volume | Type |
|------|------------|----------------|----------------|------|
| 2019 | Qiu (Qiu et al., 2019) | AS | Generic | 4 |
| 1984 | Sargent (Robert G. Sargent, 1984) | IEEE WSC | ABS | 4 |
| 2003 | Taylor (Brian J. Taylor et al., 2003) | SPIE | DL | 4 |
| 1999 | Tsai (Tsai et al., 1999) | IEEE TKDE | KBS | 4 |
| 1991 | Vinze (Vinze et al., 1991) | IM | KBS | 4 |
| 2019 | Wang (Wang et al., 2019) | ACM CHI | XAI | 4 |
| 1993 | Wells (Wells, 1993) | AAAI | KBS | 4 |
| 2018 | Zhu (J. Zhu et al., 2018) | IEEE CIG | XAI | 4 |
| 1998 | Zlatareva (N. P. Zlatareva, 1998) | DBLP | KBS | 4 |
| 2018 | Abdollahi (Abdollahi & Nasraoui, 2018) | Springer | ML | 3 |
| 1997 | Abel (Abel & Gonzalez, 1997) | FLAIRS Conference | KBS | 3 |
| 2018 | Adadi (Adadi & Berrada, 2018) | IEEE Access | XAI | 3 |
| 2018 | Agarwal (Agarwal et al., 2018) | PrePrint | Generic | 3 |
| 2016 | Amodei (Amodei et al., 2016) | PrePrint | ML | 3 |
| 2019 | Breck (Breck et al., 2019) | SysML | ML | 3 |
| 1996 | Carley (Carley, 1996) | CASOS | KBS | 3 |
| 2000 | Coenen (Coenen et al., 2000) | CUP | KBS | 3 |
| 1987 | Culbert (Culbert et al., 1987) | NASA SOAR | KBS | 3 |
| 2020 | Dağlarli (Dağlarli, 2020) | ADL | XAI | 3 |
| 1992 | Davis (Davis, 1992) | RAND | ABS | 3 |
| 2020 | Dodge (Dodge & Burnett, 2020) | ExSS-Atec | XAI | 3 |
| 2018 | Everitt (Everitt et al., 2018) | IJCAI | AGI | 3 |
| 1991 | Gilstrap (Gilstrap, 1991) | TI | KBS | 3 |
| 2019 | Glomsrud (Glomsrud et al., 2020) | ISSAV | XAI | 3 |
| 1996 | Gonzalez (Gonzalez et al., 1996) | EAAI | KBS | 3 |
| 1997 | Harmelen (Harmelen & Teije, 1997) | EUROVAV | KBS | 3 |
| 2019 | He (Y. He et al., 2020) | PrePrint | DL | 3 |
| 2020 | Heuillet (Heuillet et al., 2020) | PrePrint | RL | 3 |
| 2009 | Hibbard (Hibbard, 2009) | AGI | AGI | 3 |
| 2019 | Israelsen (Israelsen & Ahmed, 2019) | ACM CSUR | Generic | 3 |
| 2019 | Jha (Jha et al., 2019) | NeurIPS | DL | 3 |
| 2002 | Knauf (R Knauf et al., 2002) | IEEE TSMC | KBS | 3 |
| 2017 | de Laat (de Laat, 2018) | PhilosTechonol | ML | 3 |
| 1994 | Lee (S. Lee & O’Keefe, 1994) | IEEE TSMC | KBS | 3 |
| 2004 | Liu (F. Liu & Yang, 2004) | IEEE MLC | ABS | 3 |
| 1997 | Lowry (Lowry et al., 1997) | ISMIS | Generic | 3 |
| 2012 | Martinez-Balleste (Martinez-Balleste et al., 2012) | IEEE SIPC | CV | 3 |
| Year | Author(s)                                      | Journal/Conference | Paper Type | Field |
|------|-----------------------------------------------|--------------------|------------|-------|
| 2020 | Martínez-Fernandez (Martínez-Fernández et al., 2020) | PrePrint           | XAI        | 3     |
| 2017 | Mason (Mason et al., 2017a)                   | DCAART             | RL         | 3     |
| 1993 | Mengshoel (Mengshoel, 1993)                   | IEEE exp           | KBS        | 3     |
| 2005 | Menzies (Menzies & Pecheur, 2005)             | AC                 | Generic    | 3     |
| 2007 | Min (Feiyan Min et al., 2007)                 | WSC                | KBS        | 3     |
| 1997 | Murrell (Murrell & T. Plant, 1997)            | DSS                | KBS        | 3     |
| 1987 | O’Keefe (O’Keefe et al., 1987)                | IEEE exp           | KBS        | 3     |
| 2020 | Putzer (Putzer & Wozniak, 2020)               | PrePrint           | XAI        | 3     |
| 1991 | De Raedt (De Raedt et al., 1991)              | JWS                | KBS        | 3     |
| 2020 | Raji (Raji et al., 2020)                      | ACM FAT            | XAI        | 3     |
| 2004 | Sargent (Robert G. Sargent, 2004)             | IEEE WSC           | ABS        | 3     |
| 1990 | Suen (Suen et al., 1990)                      | ESA                | KBS        | 3     |
| 2019 | Sun (S. C. Sun & Guo, 2020)                   | IEEE VTC           | XAI        | 3     |
| 2006 | Yilmaz (Yilmaz, 2006)                        | CMOT               | ABS        | 3     |
| 1997 | Zaidi (Zaidi & Levis, 1997)                   | Automatica         | KBS        | 3     |
| 1996 | Abel (Abel et al., 1996)                      | FLAIRS Conference  | KBS        | 2     |
| 2016 | Aitken (Aitken, 2016)                        | PrePrint           | ABS        | 2     |
| 1998 | Antoniou (Antoniou et al., 1998)              | AI Magazine        | KBS        | 2     |
| 2019 | Arrieta (Arrieta et al., 2019)                | SciDir IF          | XAI        | 2     |
| 2018 | Bride (Bride et al., 2018)                    | ICFEM              | XAI        | 2     |
| 2020 | Dghaym (Dghaym et al., 2020)                  | AU SAV             | XAI        | 2     |
| 2015 | Dobson (Dobson, 2015)                        | JCLS               | ML         | 2     |
| 2018 | Hagras (Hagras, 2018)                        | IEEE Comp          | XAI        | 2     |
| 1999 | Hailu (Hailu & Sommer, 1999)                 | IEEE SMC           | RL         | 2     |
| 2020 | He (H. He et al., 2020)                      | IEEE IRCE          | XAI        | 2     |
| 2016 | Janssen (Janssen & Kuk, 2016)                | GIQ                | DS         | 2     |
| 2020 | Kaur (Kaur et al., 2021)                     | NBiS               | XAI        | 2     |
| 2008 | Liu (F. Liu et al., 2008)                    | IEEE SSSC          | ABS        | 2     |
| 2006 | Min (Fei-yan Min et al., 2006)               | ICMLC              | KBS        | 2     |
| 2019 | Mueller (Mueller et al., 2019)               | PrePrint           | XAI        | 2     |
| 1996 | Nourani (Nourani, 1996)                      | ACM SIGSOFT        | Generic    | 2     |
| 2020 | Pawar (Pawar et al., 2020)                   | IEEE CyberSA       | XAI        | 2     |
| 2009 | Pèpe (Pèpe et al., 2009)                    | JCG                | GA         | 2     |
| 2013 | Pitchforth (Pitchforth, 2013)                | ESA                | DL         | 2     |
| 2017 | Protiviti (Validation of Machine)            | Protiviti          | ML         | 2     |
In 2018, AI papers accounted for 3% of all peer reviewed papers published worldwide (Raymond et al., 2020). 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 twenty-fold (Raymond et al., 2020). 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 fifteen-fold starting from 2014, and the increase is highly related to the emergence of deep and reinforcement learning (Raymond et al, 2020). As the field of AI continues to grow, assurance of AI has become a more important and timely topic.
A long history of testing, validation, verification, and assurance is evident to illustrate lessons learned, pros and cons, as well as defining the future direction of AI assurance research. The next sections (4 and 5) present conclusions and recommendations for the future of AI assurance.

4. Recommendations and the future of AI assurance

4.1 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 (Heuer & Breiter 2020). Algorithmic bias for instance, whether it occurs in an unintentional or intentional manner, is found to severely limit the performance of an AI model. Given AI systems provide recommendations based on data, users’ faith in that the recommended outcomes are trustworthy, fair, and not biased is another critical challenge for AI assurance.

Applications of AI such as facial recognition using deep learning have become commonplace. Deep learning models are often exposed to adversarial inputs (such as deep-fakes), thus limiting their adoption and increasing their threat (Massoli et al., 2021). 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.

4.2 Future components of AI assurance research

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

![Figure 3: Three-dimensional AI assurance by subarea, domain, and goal](image_url)
In some cases, such as in unsupervised learning techniques, it is difficult to know what to validate or assure (Halkidi, 2001). 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 (Gao et al., 2016) and (Batarseh & Gonzalez, 2015), however, none was found that covered the three dimensional structure presented (by subarea, goal, and domain) like this review.

5. 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.

Declarations:

Ethics approval and consent to participate: Not Applicable

Consent for publication: Not Applicable

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

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

Funding: Not Applicable

Authors' contributions: FB designed the study, 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 worked on finding, arranging, and managing the papers used in the review.

Acknowledgements: Not Applicable

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

1. Abdollahi, B., & Nasraoui, O. (2018). Transparency in Fair Machine Learning: The Case of Explainable Recommender Systems. In J. Zhou & F. Chen (Eds.), Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent (pp. 21–35). Springer International Publishing. 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. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). 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]. http://arxiv.org/abs/1809.03260

6. Aitken, M. (2016). Assured Human-Autonomy Interaction through Machine Self-Confidence. University of Colorado.

7. Algorithmic Accountability Policy Toolkit. (2018). AI NOW.

8. Ali, A. L., & Schmid, F. (2014). Data Quality Assurance for Volunteered Geographic Information. In M. Duckham, E. Pebesma, K. Stewart, & A. U. Frank (Eds.), Geographic Information Science (pp. 126–141). Springer International Publishing. https://doi.org/10.1007/978-3-319-11593-1_9

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. (2016). Concrete Problems in AI Safety. ArXiv:1606.06565 [Cs]. http://arxiv.org/abs/1606.06565

11. Anderson, A., Dodge, J., Sadarangani, A., Juozapaitis, Z., Newman, E., Irvine, J., Chattopadhyay, S., Olson, M., Fern, A., & Burnett, M. (2020). Mental Models of Mere Mortals with Explanations of Reinforcement Learning. ACM Transactions on Interactive Intelligent Systems, 10(2), 1–37. https://doi.org/10.1145/3366485

12. Andert, E. P. (1992). Integrated knowledge-based system design and validation for solving problems in uncertain environments. International Journal of Man-Machine Studies, 36(2), 357–373. https://doi.org/10.1016/0020-7373(92)90023-E

13. Antoniou, G., Harmelen, F., Plant, R., & Vanthienen, J. (1998). Verification and Validation of Knowledge-Based Systems: Report on Two 1997 Events. AI Magazine, 19, 123–126.

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
15. Arifin, S. M. N., & Madey, G. R. (2015). Verification, Validation, and Replication Methods for Agent-Based Modeling and Simulation: Lessons Learned the Hard Way! In L. Yilmaz (Ed.), Concepts and Methodologies for Modeling and Simulation: A Tribute to Tuncer Ören (pp. 217–242). Springer International Publishing. https://doi.org/10.1007/978-3-319-15096-3_10

16. Arrieta, A. B., 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]. http://arxiv.org/abs/1910.10045

17. Assurance in the age of AI. (2018). EY.

18. Barr, V. B., & Klavans, J. L. (2001). Verification and validation of language processing systems: Is it evaluation? Proceedings of the Workshop on Evaluation for Language and Dialogue Systems - Volume 9, 1–7. https://doi.org/10.3115/1118053.1118058

19. Barredo-Arrrieta, A., Lana, I., & Del Ser, J. (2019). What Lies Beneath: A Note on the Explainability of Black-box Machine Learning Models for Road Traffic Forecasting. 2019 IEEE Intelligent Transportation Systems Conference (ITSC), 2232–2237. https://doi.org/10.1109/ITSC.2019.8916985

20. Batarseh, F. A., & Gonzalez, A. J. (2013). Incremental Lifecycle Validation of Knowledge-Based Systems Through CommonKADS (No. 3). 43(3), 12.

21. Batarseh, F. A., & Gonzalez, A. J. (2015). Validation of knowledge-based systems: A reassessment of the field. Artificial Intelligence Review, 43(4), 485–500. https://doi.org/10.1007/s10462-013-9396-9

22. Batarseh, A. Feras & Yang, Ruixin. (2017). Federal Data Science: Transforming Government and Agricultural Policy Using Artificial Intelligence. ISBN: 9780128124437.

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, R. K. E., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., Zhang, Y., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., & Mehta, S. (2019). AI Fairness 360: An
extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 63(4/5), 4:1-4:15. https://doi.org/10.1147/JRD.2019.2942287

27. Berndt, D. J., Fisher, J. W., Hevner, A. R., & Studnicki, J. (2001). Healthcare data warehousing and quality assurance. Computer, 34(12), 56–65. https://doi.org/10.1109/2.970578

28. Beyret, B., Shafti, A., & Faisal, A. A. (2019). Dot-to-Dot: Explainable Hierarchical Reinforcement Learning for Robotic Manipulation. 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 5014–5019. https://doi.org/10.1109/IROS40897.2019.8968488

29. Birkenbihl, C. (2020). Differences in cohort study data affect external validation of artificial intelligence models for predictive diagnostics of dementia—Lessons for translation into clinical practice. EPMA Journal, 10.

30. Bone, C., & Dragićević, S. (2010). Simulation and validation of a reinforcement learning agent-based model for multi-stakeholder forest management. Computers, Environment and Urban Systems, 34(2), 162–174. 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. (2019). Handling the adversarial attacks. Journal of Ambient Intelligence and Humanized Computing, 10(8), 2929–2943. 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]. http://arxiv.org/abs/1610.07183
39. Checco, A., Bates, J., & Demartini, G. (2020). Adversarial Attacks on Crowdsourcing Quality Control. Journal of Artificial Intelligence Research, 67, 375–408. https://doi.org/10.1613/jair.1.11332

40. Chen, H.-Y., & Lee, C.-H. (2020). Vibration Signals Analysis by Explainable Artificial Intelligence (XAI) Approach: Application on Bearing Faults Diagnosis. IEEE Access, 8, 134246–134256. https://doi.org/10.1109/ACCESS.2020.3006491

41. Chen, T., Liu, J., Xiang, Y., Niu, W., Tong, E., & Han, Z. (2019). Adversarial attack and defense in reinforcement learning from AI security view. Cybersecurity, 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., Schweitzer, 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., Ligeza, A., Laita, L., Owoc, M., Sellini, F., Spreeuwenberg, S., Vanthienen, J., Vermesan, A., & Wiratunga, N. (2000). Validation and verification of knowledge-based systems: Report on EUROVAV99. The Knowledge Engineering Review, 15(2), 187–196. https://doi.org/10.1017/S0269888900002010

45. Cohen, K. B., Hunter, L. E., & Palmer, M. (2013). Assessment of Software Testing and Quality Assurance in Natural Language Processing Applications and a Linguistically Inspired Approach to Improving It. In A. Moschitti & B. Plank (Eds.), Trustworthy Eternal Systems via Evolving Software, Data and Knowledge (pp. 77–90). Springer. https://doi.org/10.1007/978-3-642-45260-4_6

46. Cruz, F., Dazeley, R., & Vamplew, P. (2019). Memory-Based Explainable Reinforcement Learning. In J. Liu & J. Bailey (Eds.), AI 2019: Advances in Artificial Intelligence (Vol. 11919, pp. 66–77). Springer International Publishing. 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]. http://arxiv.org/abs/2006.13615

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. Dağlarlı, E. (2020). Explainable Artificial Intelligence (xAI) Approaches and Deep Meta-Learning Models. In Advances and Applications in Deep Learning (p. 18). 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]. http://arxiv.org/abs/2006.11371
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, P. B. (2018). Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability? Philosophy & Technology, 31(4), 525–541. https://doi.org/10.1007/s13347-017-0293-z
55. De Raedt, L., Sablon, G., & Bruynooghe, M. (1991). Using Interactive Concept Learning for Knowledge-base Validation and Verification. In Validation, verification and test of knowledge-based systems (pp. 177–190). John Wiley & Sons, Inc.
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., & Savat, E. (2006). A semantic validation of conceptual graphs. 13.
59. Dobson, J. (2015). Can An Algorithm Be Disturbed?: Machine Learning, Intrinsic Criticism, and the Digital Humanities. College Literature, 42, 543–564. 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. (2010). Security Assurance with Metamorphic Testing and Genetic Algorithm. 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 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, N. K., & Verheij, D. B. (2019). An Analysis of Decompositional Rule Extraction for Explainable Neural Networks. University of Groningen.

65. Edwards, D. (2000). Data Quality Assurance. In Ecological Data: Design, Management and Processing (pp. 70–91). Blackwell Science Ltd.

66. El Naqa, I., Irrer, J., Ritter, T. A., DeMarco, J., Al-Hallaq, H., Booth, J., Kim, G., Alkhatib, A., Popple, R., Perez, M., Farrey, K., & Moran, J. M. (2019). Machine learning for automated quality assurance in radiotherapy: A proof of principle using EPID data description. Medical Physics, 46(4), 1914–1921. 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]. http://arxiv.org/abs/1805.01109

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. (2016). Big Data Validation and Quality Assurance—Issues, Challenges, and Needs. 2016 IEEE Symposium on Service-Oriented System Engineering (SOSE), 433–441. https://doi.org/10.1109/SOSE.2016.63

72. Gardiner, L.-J., Carrieri, A. P., Wilshaw, J., Checkley, S., Pyzer-Knapp, E. O., & Krishna, R. (2020). Using human in vitro transcriptome analysis to build trustworthy machine learning models for prediction of animal drug toxicity. Scientific Reports, 10(1), 9522. https://doi.org/10.1038/s41598-020-66481-0

73. Gilstrap, L. (1991). Validation and verification of expert systems. Telematics and Informatics, 8(4), 439–448. 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., & Smogelü, 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, A. J., & Barr, V. (2000). Validation and verification of intelligent systems—What are they and how are they different? Journal of Experimental & Theoretical Artificial Intelligence, 12(4), 407–420. https://doi.org/10.1080/095281300454793

78. Gonzalez, A. J., Gupta, U. G., & Chianese, R. B. (1996). Performance evaluation of a large diagnostic expert system using a heuristic test case generator. Engineering Applications of Artificial Intelligence, 9(3), 275–284. 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]. http://arxiv.org/abs/1412.6572

80. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys, 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. (2020a). Explainable Artificial Intelligence for 6G: Improving Trust between Human and Machine. IEEE Communications Magazine, 58(6), 39–45. https://doi.org/10.1109/MCOM.001.2000050

83. Guo, W. (2020b). Explainable Artificial Intelligence for 6G: Improving Trust between Human and Machine. IEEE Communications Magazine, 58(6), 39–45. https://doi.org/10.1109/MCOM.001.2000050

84. Hagras, H. (2018). Toward Human-Understandable, Explainable AI. Computer, 51(9), 28–36. https://doi.org/10.1109/MC.2018.3620965

85. 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/ICSME.1999.825352

86. Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On Clustering Validation Techniques. Journal of Intelligent Information Systems, 17(2/3), 107–145.

87. Halliwell, N., & Lecue, F. (2020). Trustworthy Convolutional Neural Networks: A Gradient Penalized-based Approach. ArXiv:2009.14260 [Cs]. http://arxiv.org/abs/2009.14260

88. Han, S.-H., Kwon, M.-S., & Choi, H.-J. (2020). EXplainable AI (XAI) approach to image captioning. The Journal of Engineering, 2020(13), 589–594. https://doi.org/10.1049/joe.2019.1217

89. 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.
90. Haverinen, T. (2020). Towards Explainable Artificial Intelligence (XAI) [Master’s Thesis]. University of Jyväskylä.
91. He, C., Xing, J., Li, J., Yang, Q., Wang, R., & Zhang, X. (2015). A New Optimal Sensor Placement Strategy Based on Modified Modal Assurance Criterion and Improved Adaptive Genetic Algorithm for Structural Health Monitoring. Mathematical Problems in Engineering, 2015, 1–10. https://doi.org/10.1155/2015/626342
92. 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
93. He, Y., Meng, G., Chen, K., Hu, X., & He, J. (2020). Towards Security Threats of Deep Learning Systems: A Survey. ArXiv:1911.12562 [Cs]. http://arxiv.org/abs/1911.12562
94. Heaney, K. D., Lermusiaux, P. F. J., Duda, T. F., & Haley, P. J. (2016). Validation of genetic algorithm-based optimal sampling for ocean data assimilation. Ocean Dynamics, 66(10), 1209–1229. https://doi.org/10.1007/s10236-016-0976-5
95. 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
96. Heuillet, A., Couthouis, F., & Díaz-Rodríguez, N. (2020). Explainability in Deep Reinforcement Learning. ArXiv:2008.06693 [Cs]. http://arxiv.org/abs/2008.06693
97. Hibbard, B. (2009). Bias and No Free Lunch in Formal Measures of Intelligence. Journal of Artificial General Intelligence, 1(1), 54–61. https://doi.org/10.2478/v10229-011-0004-6
98. Huber, T. (2019). Enhancing Explainability of Deep Reinforcement Learning Through Selective Layer-Wise Relevance Propagation. 15.
99. Islam, M. A., Anderson, D. T., Pinar, A., Havens, T. C., Scott, G., & Keller, J. M. (2019). Enabling Explainable Fusion in Deep Learning with Fuzzy Integral Neural Networks. IEEE Transactions on Fuzzy Systems, 1–1. https://doi.org/10.1109/TFUZZ.2019.2917124
100. 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
101. Janssen, M., & Kuk, G. (2016). The challenges and limits of big data algorithms in technocratic governance. Government Information Quarterly, 33(3), 371–377. https://doi.org/10.1016/j.giq.2016.08.011
102. 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
103. Jiang, N., & Li, L. (2016). Doubly Robust Off-policy Value Evaluation for Reinforcement Learning. 33 Rd International Conference on Machine Learning, 48, 10.

104. 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

105. Jones, G., Willett, P., Glen, R. C., Leach, A. R., & Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. Journal of Molecular Biology, 267(3), 727–748. https://doi.org/10.1006/jmbi.1996.0897

106. Joo, H.-T., & Kim, K.-J. (2019). Visualization of Deep Reinforcement Learning using Grad-CAM: How AI Plays Atari Games? 2019 IEEE Conference on Games (CoG), 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. (2019). Trust-Based Security Mechanism for Detecting Clusters of Fake Users in Social Networks. In L. Barolli, M. Takizawa, F. Xhafa, & T. Enokido (Eds.), Web, Artificial Intelligence and Network Applications (Vol. 927, pp. 641–650). Springer International Publishing. https://doi.org/10.1007/978-3-030-15035-8_62

110. Kaur, D., Uslu, S., & Durresi, A. (2021). Requirements for Trustworthy Artificial Intelligence – A Review. In L. Barolli, K. F. Li, T. Enokido, & M. Takizawa (Eds.), Advances in Networked-Based Information Systems (Vol. 1264, pp. 105–115). Springer International Publishing. https://doi.org/10.1007/978-3-030-57811-4_11

111. Kaur, D., Uslu, S., Durresi, A., Mohler, G., & Carter, J. G. (2020). Trust-Based Human-Machine Collaboration Mechanism for Predicting Crimes. In L. Barolli, F. Amato, F. Moscato, T. Enokido, & M. Takizawa (Eds.), Advanced Information Networking and Applications (Vol. 1151, pp. 603–616). Springer International Publishing. https://doi.org/10.1007/978-3-030-44041-1_54

112. Keneni, B. M., Kaur, D., Al Bataineh, A., Devabhaktuni, V. K., Javaid, A. Y., Zaientz, J. D., & Mariner, R. P. (2019). Evolving Rule-Based Explainable Artificial Intelligence for Unmanned Aerial Vehicles. IEEE Access, 7, 17001–17016. https://doi.org/10.1109/ACCESS.2019.2893141

113. Kianifar, M. R. (2016). Application of permutation genetic algorithm for sequential model building–model validation design of experiments. Soft Comput, 20, 3023–3044. https://doi.org/DOI 10.1007/s00500-015-1929-5
114. Knauf, R., Gonzalez, A. J., & Abel, T. (2002). A framework for validation of rule-based systems. PART B, 32(3), 15.
115. Knauf, Rainer, Tsuruta, S., & Gonzalez, A. J. (2007). Toward Reducing Human Involvement in Validation of Knowledge-Based Systems. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 37(1), 120–131. https://doi.org/10.1109/TSMCA.2006.886365
116. Kohlbrenner, M., Bauer, A., Nakajima, S., Binder, A., Samek, W., & Lapuschkin, S. (2020). Towards Best Practice in Explaining Neural Network Decisions with LRP. 2020 International Joint Conference on Neural Networks (IJCNN), 1–7. https://doi.org/10.1109/IJCNN48605.2020.9206975
117. Kulkarni, A., Chong, D., & Batarseh, F. A. (2020). Foundations of data imbalance and solutions for a data democracy. In Data Democracy (pp. 83–106). Academic Press.
118. Kuppa, A., & Le-KhaC, N.-A. (2020). Black Box Attacks on Explainable Artificial Intelligence (XAI) methods in Cyber Security. 2020 International Joint Conference on Neural Networks (IJCNN), 1–8. https://doi.org/10.1109/IJCNN48605.2020.9206780
119. Kurd, Z., & Kelly, T. (2003). Safety Lifecycle for Developing Safety Critical Artificial Neural Networks. In S. Anderson, M. Felici, & B. Littlewood (Eds.), Computer Safety, Reliability, and Security (pp. 77–91). Springer. https://doi.org/10.1007/978-3-540-39878-3_7
120. Kuzlu, M., Cali, U., Sharma, V., & Guler, O. (2020). Gaining Insight Into Solar Photovoltaic Power Generation Forecasting Utilizing Explainable Artificial Intelligence Tools. IEEE Access, 8, 187814–187823. 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. (2019). Improvement in Deep Networks for Optimization Using eXplainable Artificial Intelligence. 2019 International Conference on Information and Communication Technology Convergence (ICTC), 525–530. https://doi.org/10.1109/ICTC46691.2019.8939943
122. Lee, S., & O'Keefe, R. M. (1994). Developing a strategy for expert system verification and validation. IEEE Transactions on Systems, Man, and Cybernetics, 24(4), 643–655. 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. (2018). Fair, Transparent, and Accountable Algorithmic Decision-making Processes: The Premise, the Proposed Solutions, and the Open Challenges. Philosophy & Technology, 31(4), 611–627. https://doi.org/10.1007/s13347-017-0279-x
125. Li, X.-H., Cao, C. C., Shi, Y., Bai, W., Gao, H., Qiu, L., Wang, C., Gao, Y., Zhang, S., Xue, X., & Chen, L. (2020). A Survey of Data-driven and Knowledge-aware eXplainable
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 AI 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. (1995). Knowledge validation of engineering expert systems. Advances in Engineering Software, 23(2), 97–104. https://doi.org/10.1016/0965-9978(95)00018-R

131. Lowry, M., Havelund, K., & Penix, J. (1997). Verification and validation of AI systems that control deep-space spacecraft. In Z. W. Raś & A. Skowron (Eds.), Foundations of Intelligent Systems (Vol. 1325, pp. 35–47). Springer Berlin Heidelberg. 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]. http://arxiv.org/abs/2007.15036

133. Madumal, P., Miller, T., Sonenberg, L., & Vetere, F. (2019). Explainable Reinforcement Learning Through a Causal Lens. ArXiv:1905.10958 [Cs, Stat]. http://arxiv.org/abs/1905.10958

134. Maloca, P. M., Lee, A. Y., de Carvalho, E. R., Okada, M., Fasler, K., Leung, I., Hörmann, B., Kaiser, P., Suter, S., Hasler, P. W., Zarranz-Ventura, J., Egan, C., Heeren, T. F. C., Balaskas, K., Tufail, A., & Scholl, H. P. N. (2019). Validation of automated artificial intelligence segmentation of optical coherence tomography images. PLOS ONE, 14(8), e0220063. https://doi.org/10.1371/journal.pone.0220063

135. 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

136. Marcos, M., del Pobil, A. P., & Moisan, S. (2000). Model-based verification of knowledge-based systems: A case study. IEE Proceedings - Software, 147(5), 163. https://doi.org/10.1049/ip-sen:20000896
137. 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.

138. Martínez-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

139. 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]. http://arxiv.org/abs/2003.05434

140. Martín-Guerrero, J. D., Soria-Olivas, E., Martínez-Sober, M., Climente-Martí, M., De Diego-Santos, T., & Jiménez-Torres, N. V. (2007). Validation of a Reinforcement Learning Policy for Dosage Optimization of Erythropoietin. In M. A. Orgun & J. Thornton (Eds.), AI 2007: Advances in Artificial Intelligence (pp. 732–738). Springer. https://doi.org/10.1007/978-3-540-76928-6_84

141. Mason, G., Calinescu, R., Kudenko, D., & Banks, A. (2017a). Assured Reinforcement Learning for Safety-Critical Applications.

142. Mason, G., Calinescu, R., Kudenko, D., & Banks, A. (2018). Assurance in Reinforcement Learning Using Quantitative Verification. In I. Hatzilygeroudis & V. Palade (Eds.), Advances in Hybridization of Intelligent Methods (Vol. 85, pp. 71–96). Springer International Publishing. https://doi.org/10.1007/978-3-319-66790-4_5

143. 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

144. Massoli, F. V., Carrara, F., Amato, G., & Falchi, F. (2021). Detection of Face Recognition Adversarial Attacks. Computer Vision and Image Understanding, 11.

145. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. ArXiv:1908.09635 [Cs]. http://arxiv.org/abs/1908.09635

146. 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

147. Mengshoel, O. J. (1993). Knowledge validation: Principles and practice. IEEE Expert, 8(3), 62–68. https://doi.org/10.1109/64.215224

148. Menzies, T., & Pecheur, C. (2005). Verification and Validation and Artificial Intelligence. In Advances in Computers (Vol. 65, pp. 153–201). Elsevier. https://doi.org/10.1016/S0065-2458(05)65004-8
149. 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

150. Miller, J. (1998). Active Nonlinear Test (ANTS) of Complex Simulation Models. Management Science, 44(6), Article 6.

151. Min, Feiyan, Ma, P., & Yang, M. (2007). A knowledge-based method for the validation of military simulation. 2007 Winter Simulation Conference, 1395–1402. https://doi.org/10.1109/WSC.2007.4419748

152. 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

153. 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

154. Mosqueira-Rey, E., & Moret-Bonillo, V. (2000). Validation of intelligent systems: A critical study and a tool. Expert Systems with Applications, 16.

155. Mueller, S. T., Hoffman, R. R., 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]. http://arxiv.org/abs/1902.01876

156. 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

157. Murray, B. J., Islam, M. A., Pinar, A. J., Anderson, D. T., Scott, G. J., Havens, T. C., & Keller, J. M. (2020). Explainable AI for the Choquet Integral. IEEE Transactions on Emerging Topics in Computational Intelligence, 1–10. https://doi.org/10.1109/TETCI.2020.3005682

158. Murrell, S., & T. Plant, R. (1997). A survey of tools for the validation and verification of knowledge-based systems: 1985–1995. Decision Support Systems, 21(4), 307–323. https://doi.org/10.1016/S0167-9236(97)00047-X

159. Mynuddin, M., & Gao, W. (2020). Distributed predictive cruise control based on reinforcement learning and validation on microscopic traffic simulation. IET Intelligent Transport Systems, 14(5), 270–277. https://doi.org/10.1049/iet-its.2019.0404

160. Nassar, M., Salah, K., ur Rehman, M. H., & Svetinovic, D. (2020). Blockchain for explainable and trustworthy artificial intelligence. WIREs Data Mining and Knowledge Discovery, 10(1), Article 1. https://doi.org/10.1002/widm.1340
161. 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

162. Nourani, C. F. (1996). Multi-agent object level AI validation and verification. ACM SIGSOFT Software Engineering Notes, 21(1), 70–72. https://doi.org/10.1145/381790.381802

163. O’Keefe, R. M., Balci, O., & Smith, E. P. (1987). Validating Expert System Performance. IEEE Expert, 2(4), 81–90. https://doi.org/10.1109/MEX.1987.5006538

164. On Artificial Intelligence—A European approach to excellence and trust. (2020). European Commision.

165. Onoyama, T., & Tsuruta, S. (2000). Validation method for intelligent systems. Journal of Experimental & Theoretical Artificial Intelligence, 12(4), 461–472. https://doi.org/10.1080/095281300454838

166. 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

167. Payrovnaziri, S. N., Chen, Z., Rengifo-Moreno, P., Miller, T., Bian, J., Chen, J. H., Liu, X., & He, Z. (2020). Explainable artificial intelligence models using real-world electronic health record data: A systematic scoping review. Journal of the American Medical Informatics Association, 27(7), 1173–1185. https://doi.org/10.1093/jamia/ocaa053

168. Pèpe, G., Perbost, R., Courcambeck, J., & Jouanna, P. (2009). Prediction of molecular crystal structures using a genetic algorithm: Validation by GenMolTM on energetic compounds. Journal of Crystal Growth, 311(13), 3498–3510. https://doi.org/10.1016/j.jcrysgro.2009.04.002

169. Peppler, R. A., Long, C. N., Sisterson, D. L., Turner, D. D., Bahrmann, C. P., Christensen, S. W., Doty, K. J., Eagan, R. C., Halter, T. D., Iveyh, M. D., Keck, N. N., Kehoe, K. E., Liljegren, J. C., Macduff, M. C., Mather, J. H., McCord, R. A., Monroe, J. W., Moore, S. T., Nitschke, K. L., … Wagener, R. (2008). An Overview of ARM Program Climate Research Facility Data Quality Assurance. The Open Atmospheric Science Journal, 2(1), 192–216. https://doi.org/10.2174/1874282300802010192

170. Pitchforth, J. (2013). A proposed validation framework for expert elicited Bayesian Networks. Expert Systems with Applications, 6.

171. Pocius, R., Neal, L., & Fern, A. (2019). Strategic Tasks for Explainable Reinforcement Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 33, 10007–10008. https://doi.org/10.1609/aaai.v33i01.330110007

172. Preece, A. D., Shinghal, R., & Batarekh, A. (1992). Verifying expert systems: A logical framework and a practical tool. Expert Systems with Applications, 5(3–4), 421–436. https://doi.org/10.1016/0957-4174(92)90026-O

173. 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

174. Puiutta, E., & Veith, E. M. (2020). Explainable Reinforcement Learning: A Survey. ArXiv:2005.06247 [Cs, Stat]. http://arxiv.org/abs/2005.06247

175. Putzer, H. J., & Wozniak, E. (2020). A Structured Approach to Trustworthy Autonomous/Cognitive Systems. ArXiv:2002.08210 [Cs]. http://arxiv.org/abs/2002.08210

176. Pynadath, D. V. (2018). Transparency Communication for Machine Learning in Human-Automation Interaction. In Human and Machine Learning. Springer International Publishing.

177. Qiu, S., Liu, Q., Zhou, S., & Wu, C. (2019). Review of Artificial Intelligence Adversarial Attack and Defense Technologies. Applied Sciences, 9(5), 909. https://doi.org/10.3390/app9050909

178. 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

179. Raji, I. D., Smart, A., White, R. N., 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]. http://arxiv.org/abs/2001.00973

180. 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

181. 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]. http://arxiv.org/abs/1909.09142

182. Rossi, F. (2018). Building Trust in Artificial Intelligence. Undefined. /paper/Building-Trust-in-Artificial-Intelligence-Rossi/e7a84026ac8806bd377b5b491c57096083bbb18

183. Rotman, N. H., Schapira, M., & Tamar, A. (2020). Online Safety Assurance for Deep Reinforcement Learning. ArXiv:2010.03625 [Cs]. http://arxiv.org/abs/2010.03625

184. Rovcanin, M., De Poorter, E., van den Akker, D., Moerman, I., Demeester, P., & Blondia, C. (2015). Experimental validation of a reinforcement learning based approach for a service-wise optimisation of heterogeneous wireless sensor networks. Wireless Networks, 21(3), 931–948. https://doi.org/10.1007/s11276-014-0817-8

185. 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
186. 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

187. Sargent, Robert G. (2013). Verification and validation of simulation models. Journal of Simulation, 7(1), 12–24. https://doi.org/10.1057/jos.2012.20

188. Sargent, Robert G. (1984). A tutorial on verification and validation of simulation models. Proceedings of the 16th Conference on Winter Simulation, 114–121.

189. Sargent, Robert G. (2004). Validation and Verification of Simulation Models. Proceedings of the 2004 Winter Simulation Conference, 2004., 1, 13–24. https://doi.org/10.1109/WSC.2004.1371298

190. Sargent, Robert G. (2010). Verification and validation of simulation models. Proceedings of the 2010 Winter Simulation Conference, 166–183. https://doi.org/10.1109/WSC.2010.5679166

191. 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]. http://arxiv.org/abs/1909.07082

192. Schumann, J., Gupta, P., & Liu, Y. (2010). Application of Neural Networks in High Assurance Systems: A Survey. In J. Schumann & Y. Liu (Eds.), Applications of Neural Networks in High Assurance Systems (Vol. 268, pp. 1–19). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-10690-3_1

193. Schumann, J., Gupta, P., & Nelson, S. (2003). On verification & validation of neural network based controllers.

194. Sequeira, P., & Gervasio, M. (2020). Interestingness elements for explainable reinforcement learning: Understanding agents’ capabilities and limitations. Artificial Intelligence, 288, 103367. https://doi.org/10.1016/j.artint.2020.103367

195. Sileno, G., Boer, A., & van Engers, T. (2018). The Role of Normware in Trustworthy and Explainable AI. ArXiv:1812.02471 [Cs]. http://arxiv.org/abs/1812.02471

196. Singer, E., Thurn, D. R. V., & Miller, E. R. (1995). Confidentiality Assurances and Response: A Quantitative Review of the Experimental Literature. Public Opinion Quarterly, 59(1), 66. https://doi.org/10.1086/269458

197. Sivamani, K. S., Sahay, R., & Gamal, A. E. (2020). Non-Intrusive Detection of Adversarial Deep Learning Attacks via Observer Networks. IEEE Letters of the Computer Society, 3(1), 25–28. https://doi.org/10.1109/LOCS.2020.2990897

198. 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

199. Spinner, T., Schlegel, U., Schafer, H., & El-Assady, M. (2019). explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning. IEEE
200. Srivastava, B., & Rossi, F. (2019). Towards Composable Bias Rating of AI Services. ArXiv:1808.00089 [Cs]. http://arxiv.org/abs/1808.00089

201. Stock, P., & Cisse, M. (2018). ConvNets and ImageNet Beyond Accuracy: Understanding Mistakes and Uncovering Biases. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), Computer Vision – ECCV 2018 (Vol. 11210, pp. 504–519). Springer International Publishing. https://doi.org/10.1007/978-3-030-01231-1_31

202. Suen, C. Y., Grogono, P. D., Shinghal, R., & Coallier, F. (1990). Verifying, validating, and measuring the performance of expert systems. Expert Systems with Applications, 1(2), 93–102. https://doi.org/10.1016/0957-4174(90)90019-Q

203. Sun, S. C., & 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

204. Tadj, C. (2005). Dynamic Verification of an Object-Rule Knowledge Base Using Colored Petri Nets. Systemics, Cybernetics and Informatics, 4(3), 9.

205. Tan, R., Khan, N., & Guan, L. (2020). Locality Guided Neural Networks for Explainable Artificial Intelligence. 2020 International Joint Conference on Neural Networks (IJCNN), 1–8. https://doi.org/10.1109/IJCNN48605.2020.9207559

206. Tao, C., Gao, J., & Wang, T. (2019). Testing and Quality Validation for AI Software–Perspectives, Issues, and Practices. 7, 12.

207. 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

208. Taylor, B.J., & Darrah, M. A. (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

209. 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

210. Taylor, Brian J., Darrah, M. A., & Moats, C. D. (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

211. 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
212. Thomas, J. D., & Sycara, K. (1999). The Importance of Simplicity and Validation in Genetic Programming for Data Mining in Financial Data. AAAI Technical Report, 5.

213. Tjoa, E., & Guan, C. (2020). A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 1–21. https://doi.org/10.1109/TNNLS.2020.3027314

214. Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., & Zelaya, C. G. (2020). The relationship between trust in AI and trustworthy machine learning technologies. 12.

215. Toreini, E., Aitken, M., Coopamootoo, K. P. L., Elliott, K., Zelaya, V. G., 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]. http://arxiv.org/abs/2007.08911

216. Tsai, W.-T., Vishnuvajjala, R., & Zhang, D. (1999). Verification and validation of knowledge-based systems. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 11(1), 11.

217. Turing, A. (1950). Computing Machinery and Intelligence. In Mind. Vol. 59, No. 236. pp. 433-460

218. Uslu, S., Kaur, D., Rivera, S. J., Durresi, A., & Babbar-Sebens, M. (2020a). Trust-Based Game-Theoretical Decision Making for Food-Energy-Water Management. In L. Barolli, P. Hellinckx, & T. Enokido (Eds.), Advances on Broad-Band Wireless Computing, Communication and Applications (Vol. 97, pp. 125–136). Springer International Publishing. https://doi.org/10.1007/978-3-030-33506-9_12

219. Uslu, S., Kaur, D., Rivera, S. J., Durresi, A., & Babbar-Sebens, M. (2020b). Trust-Based Decision Making for Food-Energy-Water Actors. In L. Barolli, F. Amato, F. Moscato, T. Enokido, & M. Takizawa (Eds.), Advanced Information Networking and Applications (pp. 591–602). Springer International Publishing. https://doi.org/10.1007/978-3-030-44041-1_53

220. Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. PLOS ONE, 14(11), e0224365. https://doi.org/10.1371/journal.pone.0224365

221. Validation of Machine Learning Models: Challenges and Alternatives. (2017). protiviti.

222. Varshney, K. R. (2019). Trustworthy machine learning and artificial intelligence. XRDS: Crossroads, The ACM Magazine for Students, 25(3), 26–29. https://doi.org/10.1145/3313109

223. Varshney, K. R. (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

224. 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

225. Vinze, A. S., Vogel, D. R., & Nunamaker, J. F. (1991). Performance evaluation of a knowledge-based system. Information & Management, 21(4), 225–235. https://doi.org/10.1016/0378-7206(91)90068-D

226. 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

227. Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (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

228. Wei, S., Zou, Y., Zhang, T., Zhang, X., & Wang, W. (2018). Design and Experimental Validation of a Cooperative Adaptive Cruise Control System Based on Supervised Reinforcement Learning. 22.

229. Welch, M. L., McIntosh, C., Traverso, A., Wee, L., Purdie, T. G., Dekker, A., Haibe-Kains, B., & Jaffray, D. A. (2020). External validation and transfer learning of convolutional neural networks for computed tomography dental artifact classification. Physics in Medicine & Biology, 65(3), 035017. https://doi.org/10.1088/1361-6560/ab63ba

230. Wells, S. A. (1993). The VIVA Method: A Life-cycle Independent Approach to KBS Validation. AAAI Technical Report WS-93-05, 5.

231. 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

232. 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

233. Wing, J. M. (2020). Trustworthy AI. ArXiv:2002.06276 [Cs]. http://arxiv.org/abs/2002.06276

234. Winkel, D. J. (2020). Validation of a fully automated liver segmentation algorithm using multi-scale deep reinforcement learning and comparison versus manual segmentation. European Journal of Radiology, 7.

235. 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

236. Wu, C.-H., & Lee, S.-J. (2002). KJ3—A tool assisting formal validation of knowledge-based systems. International Journal of Human-Computer Studies, 56(5), 495–524. https://doi.org/10.1006/ijhc.2002.1007
237. Xiao, Y., Pun, C.-M., & Liu, B. (2020). Adversarial example generation with adaptive gradient search for single and ensemble deep neural network. Information Sciences, 528, 147–167. https://doi.org/10.1016/j.ins.2020.04.022

238. 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

239. Yilmaz, L. (2006). Validation and verification of social processes within agent-based computational organization models. Computational and Mathematical Organization Theory, 12(4), 283–312. https://doi.org/10.1007/s10588-006-8873-y

240. 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

241. Zaidi, A. K., & Levis, A. H. (1997). Validation and verification of decision making rules. Automatica, 33(2), 155–169. https://doi.org/10.1016/S0005-1098(96)00165-3

242. Zeigler, B. P., & Nutaro, J. J. (2016). Towards a framework for more robust validation and verification of simulation models for systems of systems. The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology, 13(1), 3–16. https://doi.org/10.1177/1548512914568657

243. Zhou, J., & Chen, F. (2019). Towards Trustworthy Human-AI Teaming under Uncertainty. 5.

244. 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

245. Zhu, J., Liapis, A., Risi, S., Bidarra, R., & Youngblood, G. M. (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

246. Zlatareva, N. P. (1998). Knowledge Refinement during Developmental and Field Validation of Expert Systems. 6.

247. Zlatareva, N., & Preece, A. (1994). State of the art in automated validation of knowledge-based systems. Expert Systems with Applications, 7(2), 151–167. https://doi.org/10.1016/0957-4174(94)90034-5
**Appendix 1:** All manuscripts and their detailed scores by ranking category

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

| Year | Author     | AIs. | 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  | 0  | 0  | 0  | 0  | 1  | 0  | 1  | 0  | 0  |
| 1987 | Culbert    | KBS  | 1  | 0  | 0  | 0  | 0  | 0  | 1  | 0  | 1  | 0  |
| 1987 | O'Keefe    | KBS  | 1  | 0  | 0  | 0  | 0  | 0  | 1  | 0  | 1  | 1  |
| 1988 | Morell     | KBS  | 1  | 1  | 0  | 0  | 0  | 1  | 0  | 0  | 0  | 0  |
| 1988 | Sargent    | ABS  | 1  | 1  | 0  | 0  | 0  | 0  | 1  | 0  | 1  | 1  |
| 1989 | Becker     | KBS  | 1  | 1  | 0  | 0  | 0  | 1  | 0  | 0  | 1  | 0  |
| 1990 | Suen       | KBS  | 1  | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
| 1991 | Vinze      | KBS  | 1  | 1  | 0  | 0  | 1  | 0  | 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  | 0  | 0  | 1  | 0  | 0  | 0  | 0  | 1  |
| 1992 | Preece     | KBS  | 1  | 1  | 0  | 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  | 0  | 0  | 0  |
| 1993 | Mengshoel  | KBS  | 1  | 1  | 0  | 0  | 1  | 0  | 0  | 0  | 0  | 0  |
| 1994 | Zlatareva  | KBS  | 1  | 1  | 0  | 0  | 1  | 0  | 0  | 1  | 0  | 0  |
| 1994 | Lee        | KBS  | 1  | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  |
| 1995 | Lockwood   | KBS  | 1  | 1  | 0  | 0  | 1  | 0  | 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  |
| 1996 | Carley     | KBS  | 1  | 0  | 0  | 0  | 0  | 0  | 1  | 0  | 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 | 0 | 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 | 1 | 0 | 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 | 0 |
| 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 |
| Year | Name           | Method | 2012 | 2010 | 2009 | 2008 | 2007 | 2006 | 2005 | 2004 | 2003 | 2002 | 2001 | 2000 |
|------|----------------|--------|------|------|------|------|------|------|------|------|------|------|------|------|
| 1999 | Hailu          | RL     | 0    | 1    | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| 2000 | Mosqueira-Rey  | KBS    | 1    | 1    | 1    | 0    | 0    | 1    | 0    | 1    | 1    | 0    | 0    | 0    |
| 2000 | Marcos         | KBS    | 1    | 0    | 1    | 0    | 0    | 1    | 0    | 1    | 1    | 0    | 0    | 0    |
| 2000 | Onoyama        | KBS    | 1    | 1    | 1    | 0    | 0    | 1    | 0    | 0    | 1    | 0    | 0    | 0    |
| 2000 | Edwards        | DS     | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 1    | 1    | 0    | 0    |
| 2000 | Coenen         | KBS    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 1    | 1    | 0    | 0    |
| 2000 | Gonzalez       | Generic| 0    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 0    | 0    | 0    | 0    |
| 2001 | Berndt        | DS     | 1    | 1    | 1    | 1    | 1    | 0    | 0    | 1    | 0    | 0    | 0    | 0    |
| 2001 | Halkidi       | ML     | 1    | 1    | 1    | 0    | 0    | 1    | 0    | 0    | 1    | 1    | 0    | 0    |
| 2001 | Barr           | NLP    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 0    | 0    | 0    |
| 2002 | Wu             | KBS    | 1    | 1    | 0    | 0    | 0    | 0    | 0    | 1    | 1    | 0    | 0    | 1    |
| 2002 | Knauf          | KBS    | 1    | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 0    | 0    |
| 2003 | Schumann       | DL     | 1    | 1    | 1    | 0    | 0    | 0    | 0    | 0    | 1    | 1    | 0    | 0    |
| 2003 | Taylor         | DL     | 1    | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 0    | 1    | 0    |
| 2003 | Kurd           | DL     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 0    | 0    | 0    | 0    |
| 2004 | Liu            | ABS    | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 0    | 0    | 0    | 0    |
| 2004 | Sargent        | ABS    | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 0    | 0    | 1    | 0    |
| 2005 | Liu            | DL     | 1    | 1    | 1    | 1    | 0    | 1    | 0    | 0    | 0    | 0    | 1    | 0    |
| 2005 | Min            | KBS    | 1    | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| 2005 | Menzies        | Generic| 1    | 1    | 0    | 0    | 0    | 0    | 0    | 1    | 0    | 0    | 0    | 0    |
| 2005 | Taylor         | DL     | 1    | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| 2006 | Forster        | AGI    | 1    | 1    | 1    | 1    | 1    | 1    | 0    | 0    | 0    | 0    | 1    | 0    |
| 2006 | Taylor         | DL     | 1    | 0    | 1    | 0    | 0    | 0    | 0    | 1    | 1    | 1    | 0    | 0    |
| 2006 | Yilmaz         | ABS    | 1    | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| 2006 | Min            | KBS    | 1    | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| 2006 | Dibie-Barthélemy| KBS  | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| 2007 | Martín-Guerrero| RL     | 1    | 1    | 1    | 1    | 0    | 1    | 0    | 0    | 1    | 0    | 0    | 0    |
| 2007 | Knauf          | KBS    | 1    | 1    | 1    | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 0    |
| 2007 | Brancovici     | XAI    | 1    | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 1    |
| 2007 | Min            | KBS    | 1    | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 1    |
| 2008 | Peppler        | DS     | 1    | 1    | 1    | 1    | 1    | 1    | 0    | 1    | 0    | 0    | 0    | 0    |
| 2008 | Liu            | ABS    | 1    | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| 2009 | Tadj           | KBS    | 1    | 1    | 1    | 1    | 0    | 1    | 1    | 0    | 0    | 0    | 1    | 0    |
| 2009 | Hibbard        | AGI    | 0    | 1    | 1    | 0    | 0    | 1    | 0    | 0    | 0    | 0    | 0    | 0    |
| 2009 | Pépe           | GA     | 0    | 1    | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| 2010 | Bone           | RL     | 1    | 1    | 1    | 1    | 1    | 0    | 0    | 0    | 1    | 0    | 0    | 0    |
| 2010 | Winkler        | CV     | 1    | 1    | 1    | 0    | 0    | 1    | 0    | 0    | 1    | 1    | 0    | 0    |
| 2010 | Dong           | GA     | 1    | 1    | 1    | 0    | 0    | 1    | 0    | 0    | 0    | 0    | 1    | 0    |
| 2010 | Niazi          | ABS    | 1    | 1    | 1    | 1    | 0    | 1    | 0    | 0    | 0    | 0    | 0    | 0    |
| 2010 | Sargent        | ABS    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 0    | 0    | 1    | 0    |
| 2010 | Schumann       | DL     | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 1    |
| 2012 | Cohen          | NLP    | 0    | 1    | 1    | 1    | 0    | 1    | 0    | 0    | 0    | 0    | 0    | 0    |
| 2012 | Martinez       | CV     | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 1    | 1    | 0    |
|                 |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |
|----------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Balleste       | 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 | 0 | 0 | 1 | 0 |       |       |       |       |       |       |
|                | 2016  | Gao      | DS   | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |       |       |       |       |       |       |
|                | 2016  | Gulshan  | CV   | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |       |       |       |       |       |       |
|                | 2016  | Heaney   | GA   | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |       |       |       |       |       |       |
|                | 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  | Jilj      | ABS  | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |       |       |       |       |       |       |
|                | 2016  | Amodei    | ML   | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |       |       |       |       |       |       |
|                | 2016  | Aitken    | ABS  | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |       |       |       |       |       |       |
|                | 2016  | Janssen  | DS   | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |       |       |       |       |       |       |
|                | 2016  | Zeigler   | ABS  | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |       |       |       |       |       |       |
|                | 2016  | Wickramage | DS | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |       |       |       |       |       |       |
|                | 2017  | Liang     | DS   | 1 | 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 | 0 | 1 |       |       |       |       |       |       |
|                | 2017  | Leibovici | DS   | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |       |       |       |       |       |       |
|                | 2017  | Laat      | ML   | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |       |       |       |       |       |       |
|                | 2017  | Mason     | RL   | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |       |       |       |       |       |       |
|                | 2017  | Lepr      | ML   | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |       |       |       |       |       |       |
|                | 2018  | Wei       | RL   | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |       |       |       |       |       |       |
|                | 2018  | Alves     | ABS  | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |       |       |       |       |       |       |
|                | 2018  | Elsayer   | CV   | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 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 |       |       |       |       |       |       |
| Year | Name            | Field          | 2018 | 2019 | 2020 | 2021 | 2022 | 2023 | 2024 | 2025 |
|------|-----------------|----------------|------|------|------|------|------|------|------|------|
| 2018 | Antunes         | ML             | 1    | 1    | 1    | 0    | 0    | 1    | 0    | 0    |
| 2018 | Volz            | XAI            | 0    | 1    | 1    | 0    | 0    | 1    | 0    | 1    |
| 2018 | Al Now          | XAI            | 1    | 1    | 0    | 0    | 0    | 0    | 1    | 1    |
| 2018 | Došilović       | ML             | 1    | 0    | 0    | 0    | 0    | 0    | 1    | 1    |
| 2018 | EY              | ML             | 1    | 0    | 0    | 0    | 0    | 0    | 1    | 1    |
| 2018 | Guidotti        | XAI            | 1    | 0    | 0    | 0    | 0    | 0    | 1    | 1    |
| 2018 | Zhu             | XAI            | 1    | 0    | 0    | 0    | 0    | 0    | 1    | 1    |
| 2018 | Abdollahi       | ML             | 1    | 0    | 0    | 0    | 0    | 0    | 1    | 0    |
| 2018 | Adadi           | XAI            | 1    | 0    | 0    | 0    | 0    | 0    | 1    | 0    |
| 2018 | Agarwal         | Generic        | 1    | 1    | 0    | 0    | 0    | 0    | 0    | 0    |
| 2018 | Everitt         | AGI            | 1    | 0    | 0    | 0    | 0    | 0    | 1    | 0    |
| 2018 | Bride           | XAI            | 1    | 1    | 0    | 0    | 0    | 0    | 0    | 0    |
| 2018 | Hagras          | XAI            | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 1    |
| 2018 | Kaul            | ML             | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| 2018 | Mehri           | DL             | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| 2018 | Sileno          | XAI            | 1    | 0    | 0    | 0    | 0    | 0    | 0    | 0    |
| 2019 | Tao             | Generic        | 1    | 1    | 1    | 1    | 1    | 1    | 1    | 1    |
| 2019 | Kaur            | XAI            | 1    | 1    | 1    | 1    | 1    | 1    | 0    | 1    |
| 2019 | Batarseh        | DS             | 1    | 1    | 1    | 1    | 1    | 1    | 0    | 0    |
| 2019 | Huber           | RL             | 1    | 1    | 1    | 1    | 1    | 1    | 0    | 0    |
| 2019 | Keneni          | XAI            | 1    | 1    | 1    | 1    | 1    | 1    | 0    | 1    |
| 2019 | Maloca          | DL             | 1    | 1    | 1    | 1    | 1    | 1    | 0    | 0    |
| 2019 | Barredo-Arrieta | XAI            | 1    | 1    | 1    | 1    | 1    | 0    | 0    | 1    |
| 2019 | Chittajallu     | XAI            | 1    | 1    | 1    | 0    | 1    | 1    | 0    | 0    |
| 2019 | Ferreyra        | XAI            | 1    | 1    | 1    | 0    | 0    | 1    | 0    | 1    |
| 2019 | Lee             | XAI            | 1    | 1    | 1    | 1    | 1    | 1    | 0    | 0    |
| 2019 | Naqa            | ML             | 1    | 1    | 1    | 1    | 0    | 1    | 0    | 0    |
| 2019 | Prentzas        | XAI            | 1    | 1    | 1    | 1    | 1    | 1    | 0    | 1    |
| 2019 | Bellamy         | XAI            | 1    | 1    | 1    | 0    | 0    | 1    | 1    | 1    |
| 2019 | Beyret          | RL             | 1    | 1    | 1    | 0    | 0    | 1    | 1    | 1    |
| 2019 | Madumal         | RL             | 1    | 1    | 1    | 0    | 0    | 1    | 0    | 0    |
| 2019 | Schlegel        | XAI            | 1    | 1    | 1    | 1    | 1    | 1    | 0    | 0    |
| 2019 | Vabalas         | ML             | 1    | 1    | 1    | 1    | 0    | 1    | 0    | 0    |
| 2019 | Zhu             | RL             | 1    | 1    | 1    | 0    | 0    | 1    | 1    | 0    |
| 2019 | Chen            | RL             | 1    | 1    | 0    | 0    | 0    | 0    | 1    | 1    |
| 2019 | Cruz            | RL             | 1    | 1    | 1    | 0    | 0    | 1    | 0    | 0    |
| 2019 | Dupuis          | XAI            | 1    | 1    | 1    | 0    | 0    | 1    | 0    | 0    |
| 2019 | Joo             | RL             | 1    | 1    | 1    | 0    | 0    | 1    | 1    | 0    |
| 2019 | Ren             | DL             | 0    | 1    | 1    | 0    | 0    | 1    | 1    | 0    |
| 2019 | Srivastava      | NLP            | 1    | 1    | 1    | 0    | 0    | 1    | 0    | 0    |
| 2019 | Uslu            | XAI            | 1    | 1    | 1    | 0    | 0    | 1    | 0    | 0    |
| 2019 | Yoon            | XAI            | 1    | 1    | 1    | 0    | 0    | 1    | 0    | 0    |
| 2019 | Zhou            | ML             | 1    | 1    | 1    | 0    | 0    | 1    | 0    | 0    |
| Name          | Type | ML | DL | XAI | DS | RL | DS |
|---------------|------|----|----|-----|----|----|----|
| 2019 Mehrabi  | ML   | 1  | 0  | 0   | 0  | 0  | 0  |
| 2019 Meskauskas | XAI  | 1  | 1  | 1   | 0  | 0  | 1  |
| 2019 Nassar   | XAI  | 1  | 1  | 0   | 0  | 0  | 0  |
| 2019 Qiu      | Generic | 1 | 0  | 0   | 0  | 0  | 1  |
| 2019 Wang     | XAI  | 1  | 1  | 0   | 0  | 0  | 0  |
| 2019 Breck    | ML   | 1  | 1  | 0   | 0  | 0  | 0  |
| 2019 Glomrsud | XAI  | 1  | 0  | 0   | 0  | 0  | 0  |
| 2019 He       | DL   | 1  | 0  | 0   | 0  | 0  | 0  |
| 2019 Israelson| Generic | 1 | 0  | 0   | 0  | 0  | 0  |
| 2019 Jha      | DL   | 1  | 0  | 1   | 0  | 0  | 0  |
| 2019 Sun      | XAI  | 0  | 1  | 1   | 0  | 0  | 0  |
| 2019 Dghayym  | XAI  | 0  | 0  | 0   | 0  | 0  | 0  |
| 2019 Mueller  | XAI  | 1  | 0  | 0   | 0  | 0  | 0  |
| 2019 Protiviti| ML   | 0  | 0  | 0   | 0  | 0  | 0  |
| 2019 Spada    | XAI  | 0  | 1  | 0   | 0  | 0  | 0  |
| 2019 Pocius   | RL   | 1  | 0  | 0   | 0  | 0  | 0  |
| 2019 Rossi    | XAI  | 1  | 0  | 0   | 0  | 0  | 0  |
| 2019 Varshney | ML   | 0  | 0  | 0   | 0  | 0  | 0  |
| 2020 D'Alterio| XAI  | 1  | 1  | 1   | 1  | 1  | 1  |
| 2020 Anderson | RL   | 1  | 1  | 1   | 1  | 1  | 0  |
| 2020 Birkenbihl| ML  | 1  | 1  | 1   | 1  | 1  | 0  |
| 2020 Checco   | DS   | 1  | 1  | 1   | 1  | 1  | 0  |
| 2020 Chen     | XAI  | 1  | 1  | 1   | 1  | 1  | 0  |
| 2020 EASA     | DL   | 1  | 1  | 1   | 1  | 1  | 0  |
| 2020 Kulkarni | DS   | 1  | 1  | 1   | 1  | 1  | 0  |
| 2020 Kuppa    | XAI  | 1  | 1  | 1   | 1  | 1  | 0  |
| 2020 Kuzlu    | XAI  | 1  | 1  | 1   | 1  | 1  | 0  |
| 2020 Spinner  | XAI  | 1  | 1  | 1   | 1  | 1  | 0  |
| 2020 Winkel   | RL   | 1  | 1  | 1   | 1  | 1  | 0  |
| 2020 Gardiner | ML   | 1  | 1  | 1   | 1  | 1  | 0  |
| 2020 Guo      | XAI  | 1  | 1  | 1   | 1  | 1  | 0  |
| 2020 Han      | XAI  | 1  | 1  | 1   | 1  | 1  | 0  |
| 2020 Kohlbrenner| XAI  | 1  | 1  | 1   | 1  | 1  | 0  |
| 2020 Malolan  | XAI  | 1  | 1  | 1   | 1  | 1  | 0  |
| 2020 Payrovni| ML   | 1  | 1  | 0   | 1  | 1  | 0  |
| 2020 Sequeira | RL   | 1  | 1  | 1   | 1  | 1  | 0  |
| 2020 Sivamani | DL   | 1  | 1  | 1   | 1  | 1  | 0  |
| 2020 Tan      | XAI  | 1  | 1  | 1   | 1  | 1  | 0  |
| 2020 Tao      | XAI  | 1  | 1  | 1   | 1  | 1  | 0  |
| 2020 Welch    | DL   | 1  | 1  | 1   | 1  | 1  | 0  |
| 2020 Xiao     | DL   | 1  | 1  | 1   | 1  | 1  | 0  |
| 2020 Halliwell| DL   | 1  | 1  | 1   | 1  | 1  | 0  |
| 2020 Heuer    | ML   | 1  | 1  | 1   | 1  | 1  | 0  |
| Year | Name            | Area | 2020 | 2021 |
|------|-----------------|------|------|------|
| 2020 | Kaur            | XAI  | 1    | 1    |
| 2020 | Mackowiak      | CV   | 1    | 1    |
| 2020 | Ragot           | ML   | 1    | 1    |
| 2020 | Rotman          | RL   | 1    | 1    |
| 2020 | Sarathy         | XAI  | 1    | 1    |
| 2020 | Uslu            | XAI  | 0    | 1    |
| 2020 | Cruz            | RL   | 1    | 1    |
| 2020 | He              | RL   | 1    | 0    |
| 2020 | Islam           | XAI  | 0    | 1    |
| 2020 | Mynuddin        | RL   | 1    | 0    |
| 2020 | Puiutta         | RL   | 1    | 1    |
| 2020 | Toreini         | ML   | 1    | 0    |
| 2020 | Toreini         | ML   | 1    | 0    |
| 2020 | Diallo          | XAI  | 1    | 0    |
| 2020 | Guo             | XAI  | 1    | 0    |
| 2020 | Haverinen       | XAI  | 0    | 1    |
| 2020 | Katell          | XAI  | 1    | 0    |
| 2020 | Murray          | XAI  | 0    | 1    |
| 2020 | Taylor          | XAI  | 1    | 1    |
| 2020 | Tjoa            | ML   | 1    | 0    |
| 2020 | Varshney        | ML   | 1    | 1    |
| 2020 | Wieringa        | XAI  | 1    | 0    |
| 2020 | Wing            | ML   | 1    | 0    |
| 2020 | Das             | XAI  | 1    | 0    |
| 2020 | Li              | XAI  | 0    | 0    |
| 2020 | Dağlarli        | XAI  | 1    | 0    |
| 2020 | Dodge           | XAI  | 1    | 1    |
| 2020 | Heuillet        | RL   | 1    | 0    |
| 2020 | Martinez-        | XAI  | 1    | 1    |
| 2020 | Fernandez       |      |      |      |
| 2020 | Putzer          | XAI  | 0    | 1    |
| 2020 | Raji            | XAI  | 1    | 0    |
| 2020 | Arrieta         | XAI  | 1    | 0    |
| 2020 | He              | XAI  | 1    | 0    |
| 2020 | Kaur            | XAI  | 1    | 0    |
| 2020 | Pawar           | XAI  | 0    | 1    |
| 2020 | Brennen         | XAI  | 0    | 0    |
| 2020 | European        | XAI  | 0    | 0    |
| 2020 | Commission      |      |      |      |
| 2021 | Massoli         | DL   | 1    | 1    |